



500.43154X00

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant(s): Y. OGAWA, et al.

Serial No.: 10/671,718

Filed: September 29, 2003

Title: METHOD AND DEVICE FOR RELEVANT DOCUMENT SEARCH

LETTER CLAIMING RIGHT OF PRIORITY

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

November 6, 2003

Sir:

Under the provisions of 35 USC 119 and 37 CFR 1.55, the applicant(s) hereby
claim(s) the right of priority based on:

Japanese Patent Application No. 2003-089633
Filed: March 28, 2003

A certified copy of said Japanese Patent Application is attached.

Respectfully submitted,

ANTONELLI, TERRY, STOUT & KRAUS, LLP

Carl I. Brundidge

Registration No.: 29,621

CIB/rr
Attachment

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 2 0 0 3 年 3 月 2 8 日
Date of Application:

出 願 番 号 特 願 2 0 0 3 - 0 8 9 6 3 3
Application Number:
[ST. 10/C]: [J P 2 0 0 3 - 0 8 9 6 3 3]

出 願 人 株式会社日立製作所
Applicant(s):

2 0 0 3 年 1 0 月 2 日

特許庁長官
Commissioner,
Japan Patent Office

今 井 康 夫



出証番号 出証特 2 0 0 3 - 3 0 8 1 1 9 9

【書類名】 特許願

【整理番号】 K02011981A

【あて先】 特許庁長官殿

【国際特許分類】 G06F 17/30

【発明者】

【住所又は居所】 神奈川県川崎市幸区鹿島田 8 9 0 番地 株式会社日立製作所 ビジネスソリューション事業部内

【氏名】 小川 祐一

【発明者】

【住所又は居所】 神奈川県川崎市幸区鹿島田 8 9 0 番地 株式会社日立製作所 ビジネスソリューション事業部内

【氏名】 松林 忠孝

【発明者】

【住所又は居所】 神奈川県横浜市戸塚区戸塚町 5 0 3 0 番地 株式会社日立製作所 ソフトウェア事業部内

【氏名】 山本 伸也

【特許出願人】

【識別番号】 000005108

【氏名又は名称】 株式会社日立製作所

【代理人】

【識別番号】 100075096

【弁理士】

【氏名又は名称】 作田 康夫

【手数料の表示】

【予納台帳番号】 013088

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1
【プルーフの要否】 要

【書類名】 明細書

【発明の名称】

類似文書検索方法および類似文書検索装置

【特許請求の範囲】

【請求項 1】

検索対象の文書（以下、対象文書という）の中から文書を検索する類似文書検索方法であって、

検索条件として入力された全文検索条件を取得し、

前記対象文書を複数の部分に分割し、

該分割した前記対象文書の各部分に対して前記全文検索条件に対する類似度を算出し、

該算出した類似度と予め定められた閾値とを比較して、前記分割された各部分が前記全文検索条件に適合している部分であるか否かを判定し、

該判定結果をもとに、前記分割された部分を含む対象文書の前記全文検索条件に対する詳細度を算出することを特徴とする類似文書検索方法。

【請求項 2】

前記類似文書検索方法はさらに、

前記対象文書の前記全文検索条件との類似度を算出し、

該算出した前記全文検索条件に対する類似度と、前記算出した前記対象文書の前記全文検索条件に対する詳細度とを表示することを特徴とする請求項 1 記載の類似文書検索方法。

【請求項 3】

検索対象の文書（以下、対象文書という）の中から関連する文書を検索する類似文書検索装置において、

検索条件である種文書を取得する種文書取得手段と、

前記対象文書を複数の部分に分割する分割手段と、

前記取得された種文書をもとに、該分割された前記対象文書の各部分について前記種文書に対する類似度を算出する類似度算出手段と、

該算出された部分ごとの類似度が、所定の値を超えているか否かを判断して、

前記対象文書の前記種文書に対する詳細度を算出する詳細度算出手段とを備えることを特徴とする類似文書検索装置。

【請求項 4】

前記類似文書検索装置はさらに、

前記対象文書に対する全文検索における全文検索条件を解析する全文検索条件解析手段と、

該解析された全文検索条件をもとに、前記分割された各部分に対して全文検索を行い、前記各部分の前記全文検索条件に対する適合度を算出する全文検索条件適合度算出手段とを備え、

前記詳細度算出手段は、該全文検索条件適合度と、前記類似度算出手段が算出した前記分割された各部分について前記種文書に対する類似度とを用いて、前記対象文書の前記種文書に対する詳細度を算出することを特徴とする請求項 3 記載の類似文書検索装置。

【請求項 5】

前記類似文書検索装置はさらに、

前記種文書に対する詳細度または前記種文書に対する類似度をキーとして、検索対象である複数の対象文書に対して順位をつけて表示する表示手段を備えることを特徴とする請求項 3 記載の類似文書検索装置。

【請求項 6】

検索対象の文書（以下、対象文書という）の中から関連する文書を検索する類似文書検索プログラムを格納した記憶媒体であって、

前記対象文書を検索する検索条件である種文書を取得するステップと、

前記対象文書を複数の部分に分割するステップと、

前記取得された種文書をもとに、前記分割された各部分の前記種文書に対する類似度を算出するステップと、

該算出された類似度を、所定の閾値と比較するステップと、

該比較の結果を用いて、前記分割された各部分が前記種文書に適合しているか否かを判定して、前記種文書に適合している部分の数を集計するステップと、

該集計した数をもとに、前記対象文書の前記種文書に対する詳細度を算出する

ステップとを備えることを特徴とする類似文書検索プログラムを格納した記憶媒体。

【請求項 7】

予め記憶された検索対象の文書（以下、対象文書という）と、検索条件である文書（以下、種文書という）との関連性を判定する文書間関連性判定方法であって、

前記対象文書を複数の部分に分割し、

該分割した前記対象文書の各部分に対して前記種文書に対する類似度を算出し

、

該算出した類似度と予め定められた閾値とを比較して、前記分割された各部分が前記種文書に適合している部分であるか否かを判定し、

該判定結果をもとに、前記種文書に適合している部分の数を集計し、

該集計した部分の数をもとに、前記分割された部分を含む対象文書の前記種文書に対する詳細度を算出することを特徴とする文書間関連性判定方法。

【請求項 8】

前記文書間関連性判定方法はさらに、

前記対象文書の前記種文書に対する類似度を算出し、

該算出した前記種文書に対する対象文書の類似度と、前記算出した前記種文書に対する対象文書の詳細度の少なくとも一方を出力することを特徴とする請求項 7 記載の文書間関連性判定方法。

【請求項 9】

前記文書間関連性判定方法はさらに、

前記対象文書に対して全文検索を行う場合の全文検索条件を取得し、

該取得した全文検索条件を用いて、前記分割された対象文書の各部分に対して全文検索を行い、前記分割された対象文書の部分ごとに前記全文検索条件に対する適合度を算出し、

該算出した前記全文検索条件に対する適合度と、前記算出した対象文書の部分ごとの前記種文書に対する類似度とを用いて、対象文書の中で前記検索に適合する部分を検出することを特徴とする請求項 7 記載の文書間関連性判定方法。

【発明の詳細な説明】**【0 0 0 1】****【発明の属する技術分野】**

本発明は、ユーザが指定した文書間の類似性の判定指標を算出する文書間関連度算出方法とこれを用いた類似文書検索方法に関する。

【0 0 0 2】**【従来技術】**

近年、パーソナルコンピュータやインターネットの普及に伴い、電子化された文書が大量に存在するようになった。その大量の文書の中からユーザーが目的とする文書を効率よく検索する文書検索技術が盛んに開発されており、中でも検索条件として入力された文書（以下、種文書と呼ぶ）と類似した文書を検索する類似文書検索が注目されている。

【0 0 0 3】

この類似文書検索に関して、特開平 9 - 1 6 0 9 2 8 号公報には、種文書を構成する文と、種文書に対する類似度を算出する文書（以下、対象文書と呼ぶ）を構成する文の全組み合わせに対して文間の類似度を算出し、それらの類似度を加算することで文書全体の類似度を算出する技術が開示されている。例えば、種文書が A、B の 2 文で構成され、対象文書が C、D、E の 3 文で構成されている場合、種文書に関する対象文書の類似度は、（A と C の類似度）、（A と D の類似度）、（A と E の類似度）、（B と C の類似度）、（B と D の類似度）、（B と E の類似度）の和として算出される。これにより、種文書に関する内容が対象文書の全体で類似している場合に高い類似度の値が算出される。

【0 0 0 4】

【特許文献 1】 特開平 9 - 1 6 0 9 2 8 号公報

【発明が解決しようとする課題】

しかし、上記従来技術では、ある文間の類似度が極端に高い場合、他の文間の類似度が低くても文書全体の類似度としては高くなってしまう場合がある。すなわち、ある対象文書に対して高い類似度が算出された場合、対象文書の全体が類似している場合と対象文書の一部が類似している場合が考えられる。検索者はこ

これらの違いを区別できないため、ユーザは目的に応じた種文書に関する効率的な検索が行なえない。例えば、種文書に記載された内容に関して幅広く情報を得るために文書全体で類似している対象文書を参照したい場合、上記従来技術を用いて算出された類似度では判断できない。

【0 0 0 5】

本発明の目的は、文書の類似性を判断するための指標を提示する類似文書検索方法を提供することにある。

【0 0 0 6】

【課題を解決するための手段】

上記目的を達成するために本発明は、予め記憶された検索対象文書の中から文書を検索する検索条件として入力された種文書に含まれる文字列を抽出し、対象文書を複数の部分に分割して、分割した対象文書の各部分に含まれる文字列を抽出し、これら文字列を比較して、前記分割された部分ごとに前記種文書に対する類似度を算出するとともに、その類似度と予め定められた閾値とを比較して、分割された各部分が種文書に適合している部分であるか否かの判定結果をもとに、対象文書の前記種文書に対する詳細度を算出する構成を採用した。

【0 0 0 7】

【発明の実施の形態】

以下に、本発明の第一の実施例について説明する。

【0 0 0 8】

図 1 は、本実施例で示す文書検索システムの全体構成図を示す。本システムは、ディスプレイ 1 0 0、キーボード 1 0 1、中央演算処理装置（CPU） 1 0 2、磁気ディスク装置 1 0 3、フレキシブルディスクドライブ（FDD） 1 0 4、主メモリ 1 0 5、これらを結ぶバス 1 0 6 および他の機器と本システムを接続するネットワーク 1 0 7 から構成される。

【0 0 0 9】

磁気ディスク装置 1 0 3 は二次記憶装置の一つであり、テキスト 1 7 0 が格納される。FDD 1 0 4 を介してフレキシブルディスク 1 0 8 に格納されている情報が、主メモリ 1 0 5 あるいは磁気ディスク装置 1 0 3 へ読み込まれる。

【0010】

主メモリ 105 には、システム制御プログラム 110、登録制御プログラム 111、検索制御プログラム 112、文書ファイル取得プログラム 120、テキスト登録プログラム 121、種文書解析プログラム 130、テキスト読込プログラム 131、類似度算出プログラム 132、詳細度算出制御プログラム 133、ブロック分割プログラム 140、ブロック別類似度算出プログラム 141、詳細度算出プログラム 142、結果出力プログラム 134 及び共有ライブラリ 150 が記憶され、またワークエリア 160 が確保される。なお、共有ライブラリ 150 は、特徴語抽出プログラム 151 で構成される。

【0011】

システム制御プログラム 110 は、登録制御プログラム 111 および検索制御プログラム 112 で構成される。登録制御プログラム 111 は、文書ファイル取得プログラム 120 およびテキスト登録プログラム 121 で構成される。検索制御プログラム 112 は、種文書解析プログラム 130、テキスト読込プログラム 131、類似度算出プログラム 132、詳細度算出制御プログラム 133 および結果出力プログラム 134 で構成されるとともに、特徴語抽出プログラム 151 を呼び出す構成をとる。詳細度算出制御プログラム 133 は、ブロック分割プログラム 140、ブロック別類似度算出プログラム 141 および詳細度算出プログラム 142 で構成されるとともに、特徴語抽出プログラム 151 を呼び出す構成をとる。

【0012】

登録制御プログラム 111 および検索制御プログラム 112 は、ユーザによるキーボード 101 からの入力に応じてシステム制御プログラム 110 によって起動される。登録制御プログラム 111 は、文書ファイル取得プログラム 120 とテキスト登録プログラム 121 を制御する。検索制御プログラム 112 は、種文書解析プログラム 130、特徴語抽出プログラム 151、テキスト読込プログラム 131、類似度算出プログラム 132、詳細度算出制御プログラム 133 および結果出力プログラム 134 を制御する。

【0013】

本実施例では、キーボード 101 から入力されたコマンドにより登録制御プログラム 111 および検索制御プログラム 112 が起動されるものとしたが、他の入力装置を介して入力されたコマンドあるいはイベントにより起動されるものであってもかまわない。

【0014】

また、これらのプログラムを磁気ディスク 103、フレキシブルディスク 108、MO、CD-ROM、DVD等の記憶媒体（図 1 には示していない）に格納し、駆動装置を介して主メモリ 105 に読み込み、CPU 102 によって実行することが可能である。また、これらのプログラムをネットワーク 107 を介して主メモリ 105 に読み込み、CPU 102 によって実行することも可能である。

【0015】

また、本実施例ではテキスト 170 は磁気ディスク装置 103 に格納されるものとしたが、フレキシブルディスク 108、MO、CD-ROM、DVD等の記憶媒体（図 1 には示していない）に格納し、駆動装置を介して主メモリ 105 に読み込み利用することも可能であるし、あるいはネットワーク 107 を介して、他のシステムに接続された記憶媒体（図 1 には示していない）に格納されるものとしてもよい。また、さらにはネットワーク 107 に直接接続された記憶媒体に格納されるものとしても構わない。

【0016】

次に、システム制御プログラム 110 の処理手順について説明する。システム制御プログラム 110 は、まずキーボード 101 から入力されたコマンドを解析する。この結果が登録実行のコマンドであると解析された場合には、登録制御プログラム 111 を起動して、文書の登録を行う。また、検索実行のコマンドであると解析された場合には、検索制御プログラム 112 を起動して、検索条件として入力された複数の単語や文、文章あるいは文書（以下、まとめて種文書と呼ぶ）に関連した内容を含む文書の検索を行う。

【0017】

次に、システム制御プログラム 110 により起動される登録制御プログラム 111 の処理手順について説明する。登録制御プログラム 111 は、まず文書ファ

イル取得プログラム 1 2 0 を起動し、FDD 1 0 4 を介してフレキシブルディスク 1 0 8 に格納されている文書ファイルを読み込む。次に、テキスト登録プログラム 1 2 1 を起動して、前記文書ファイル取得プログラム 1 2 0 で読み込まれた文書ファイルからテキストを抽出し、磁気ディスク装置 1 0 3 にテキスト 1 7 0 として格納する。

【0 0 1 8】

なお、文書ファイルはフレキシブルディスク 1 0 8 に格納されているものとしたが、MO、CD-ROM、DVD等の記憶媒体（図 1 には示していない）に格納されるものとしてもよいし、ネットワーク 1 0 7 を介して、他のシステムに接続された記憶媒体（図 1 には示していない）に格納されるものとしてもよい。また、文書ファイル取得プログラム 1 2 0 で読み込まれた文書ファイルはテキストが抽出できるものならばよく、テキストファイルとして保存されているものであってもよいし、アプリケーションソフトの保存形式であってもよい。

【0 0 1 9】

システム制御プログラム 1 1 0 により起動される検索制御プログラム 1 1 2 の処理手順について図 2 を用いて説明する。検索制御プログラム 1 1 2 は、まず種文書解析プログラム 1 3 0 を起動し、検索条件で指定された種文書を読み込み、ワークエリア 1 6 0 に格納する（ステップ 2 0 0）。次に、特徴語抽出プログラム 1 5 1 を起動し、前記種文書解析プログラム 1 3 0 によりワークエリア 1 6 0 に格納された種文書から自立した意味を持つ文字列（以下、特徴語と呼ぶ）を抽出し、ワークエリア 1 6 0 に格納する（ステップ 2 1 0）。

【0 0 2 0】

テキスト 1 7 0 に含まれるすべてのテキストに対して、ステップ 2 2 1 ～ステップ 2 2 3 を繰り返し実行する（ステップ 2 2 0）。まず、テキスト読込プログラム 1 3 1 を起動し、磁気ディスク装置 1 0 3 に格納されているテキスト 1 7 0 からテキストを 1 つ読み込む（ステップ 2 2 1）。次に、類似度算出プログラム 1 3 2 を起動し、前記テキスト読込プログラム 1 3 1 により読み込まれたテキストに対し、一般的な類似文書検索技術を用いて種文書に対するテキストの類似度を算出し、ワークエリア 1 6 0 に格納する（ステップ 2 2 2）。次に、詳細度算

出制御プログラム 133 を起動し、前記テキスト読込プログラム 131 により読み込まれたテキスト全体に対し、種文書に関する内容が占める割合（以下、詳細度と呼ぶ）を算出し、ワークエリア 160 に格納する（ステップ 223）。

【0021】

そして、結果出力プログラム 134 を起動し、前記類似度算出プログラム 132 により算出された類似度と前記詳細度算出制御プログラム 133 により算出された詳細度を各テキストに対して出力する（ステップ 230）。

【0022】

なお、特徴語抽出プログラム 151 により抽出される特徴語は、漢字やカタカナといった文字種間や文章中に存在するスペースなどの区切り文字により分割された文字列であってもよいし、形態素解析により抽出される単語や n-gram として抽出される文字列であってもよいし、その他の方法により抽出された文字列であってもかまわない。

【0023】

ステップ 222 における類似度算出処理は、上記従来技術に記載した類似度算出方法や、ベクトル空間法における余弦尺度を用いた類似度算出方法などを適用することができる。

【0024】

また、類似度および詳細度が算出されるテキスト 170 は、磁気ディスク装置 103 に格納されるものとしたが、フレキシブルディスク 108、MO、CD-ROM、DVD 等の記憶媒体（図 1 には示していない）に格納されるものとして、あるいはネットワーク 107 を介して、他のシステムに接続された記憶媒体（図 1 には示していない）に格納されるものとしてもよい。

【0025】

前記ステップ 220 ではテキスト 170 に含まれるすべてのテキストに対して、ステップ 221～ステップ 225 を繰り返すものとしたが、テキスト 170 に含まれる一部のテキストに対して繰り返されるものであってもよい。

【0026】

本実施例ではテキスト読込プログラム 131 によって読み込まれたテキスト全

体に対して類似度および詳細度を算出するものとしたが、テキスト全体でなくともよく、テキストの一部を対象に本発明を適用することが可能である。

【0027】

次に、検索制御プログラム112により起動される詳細度算出制御プログラム133の処理手順（図2のステップ223の詳細）について、図3に示すPAD図を用いて説明する。

【0028】

まず、種文書に適合しているブロックの数（以下、適合ブロック数と呼ぶ）とテキストに含まれるブロックの数（以下、総ブロック数と呼ぶ）の初期値をともに0と設定する（ステップ300）。次に、ブロック分割プログラム140を起動し、前記テキスト読込プログラム131で読み込まれたテキストを文、段落、章などの部分（以下、これらをまとめてブロックと呼ぶ）に分割する（ステップ310）。

【0029】

前記ステップ310で分割された各ブロックに対して、それぞれステップ321～ステップ325を繰り返し実行する（ステップ320）。まず、特徴語抽出プログラム151を起動し、ステップ310で分割された各ブロックから特徴語を抽出する（ステップ321）。次に、ブロック別類似度算出プログラム141を起動し、図2のステップ210で抽出された種文書の特徴語と、ステップ321で抽出された各ブロックの特徴語から、種文書に対する各ブロックの類似度を式1を用いて算出する（ステップ322）。

【0030】

【式1】

次に、ステップ322で算出されたブロックの類似度を、種文書に対する適合性を判定する際の基準値（以下、種文書適合性判定閾値と呼ぶ）と比較する（ステップ323）。この結果、ブロックの類似度が種文書適合性判定閾値以上であった場合、そのブロックを種文書に適合しているブロック（以下、適合ブロック）と判定し、適合ブロック数を1加算するとともに（ステップ324）、総ブロック数を1加算する（ステップ325）。ステップ323において、ブロックの

類似度が閾値以下であった場合は、適合ブロック数は 1 加算されず、総ブロック数のみが 1 加算される（ステップ 3 2 5）。

【 0 0 3 1 】

ステップ 3 1 0 で分割されたすべてのブロックに対して、ステップ 3 2 1 ～ 3 2 5 の処理を終了したら、詳細度算出プログラム 1 4 2 を起動し、ステップ 3 2 4 およびステップ 3 2 5 で計数された適合ブロック数と総ブロック数から、式 2 を用いて種文書に対する該テキストの詳細度を算出する（ステップ 3 3 0）。

【 0 0 3 2 】

【式 2】

最後に、ステップ 3 3 0 で算出された種文書に対する該テキストの詳細度をワークエリア 1 6 0 に格納する（ステップ 3 4 0）。

【 0 0 3 3 】

なお、上記ステップ 3 2 2 におけるブロックの類似度の算出には、式 1 に示した類似度算出式を適用したが、ベクトル空間法における余弦尺度など他の類似度算出式を適用してもよい。

【 0 0 3 4 】

次に、本実施例における文書検索システムの検索処理の流れについて、図 4 および図 5 を用いて説明する。

【 0 0 3 5 】

図 4 に示した例は、文書 1 「In The Sports Championship Cup, Country-A broke through the primary league for the first time. Country-A played a match against Country-B of the Championship ranking highest in H group at the first game, and though troubled, and was a draw. Then, both the Country-C game and the Country-D game gained a victory with offensive strategy, and passed the brilliant H group by the 1st place. A final tournament is due to play a match against Country-E.」 および文書 2 「Country-A is still in the state of economic depression. If there is bright news that induces an economic big effect, can Country-A escape from economic depression? The Sports Championship Cup was held for the first time in Countr

y-A, and Country-A passed H group including Country-B, Country-C, and Country-D by the 1st place on the other day. However, it was not able to become an explosive to economic recovery and an economic big effect could not be acquired.」 (文書 2 は図 4 に示していない) が磁気ディスク装置 1 0 3 に格納された類似文書検索システムにおいて、種文書として「The Sports Championship Cup held for the first time in Country-A , and Country-A passed H group including Country-B, Country-C, and Country-D by the 1st place .」 が入力された場合の例を示している。なお、本図は、種文書解析プログラム 1 3 0 により、検索条件として入力された種文書が文書 4 0 0 として読み込まれ、テキスト読込プログラム 1 3 1 により、文書 1 がテキスト 4 1 0 として読み込まれた状態である。

【 0 0 3 6 】

まず、類似度算出プログラム 1 3 2 が実行され、前記テキスト読込プログラム 1 3 1 により読み込まれたテキスト 4 1 0 と前記種文書解析プログラム 1 3 0 により読み込まれた種文書 4 0 0 から、種文書に対するテキスト 4 1 0 の類似度を算出する (図 2 のステップ 2 2 2) 。本実施例では、類似度を上記従来技術に記載された技術を適用して算出し、類似度算出結果 4 2 0 として類似度が “1.06” と算出され、ワークエリア 1 6 0 に格納される。ここで、種文書に含まれる文の重みはすべて “1” とする。

【 0 0 3 7 】

次に、ブロック分割プログラム 1 4 0 が実行され、テキスト 4 1 0 をブロック単位へ分割する (図 3 のステップ 3 1 0) 。本図に示した例では、テキスト 4 1 0 に対し “.” (ピリオド) を区切り文字としてブロック単位に分割しており、この結果としてブロック分割結果 4 3 0 が出力されている。本図に示したブロック分割結果 4 3 0 は、ブロック 1 「In The Sports Championship Cup, Country-A broke through the primary league for the first time.」、ブロック 2 「Country-A played a match against Country-B of the Championship ranking highest in H group at the first game, and though troubled, and was a draw.」、ブロック 3 「Then, both the Country-C game and the Country-D game gai

ned a victory with offensive strategy, and passed the brilliant H group by the 1st place.」およびブロック 4 「A final tournament is due to play a match against Country-E.」であり、これらのブロックがワークエリア 1 6 0 に格納されている。

【0 0 3 8】

一方、特徴語抽出プログラム 1 5 1 が実行され、前記種文書解析プログラム 1 3 0 により読み込まれた種文書 4 0 0 から “Sports”、“Championship”、“Cup”、“held”、“first”、“time”、“Country-A”、“passed”、“group”、“including”、“Country-B”、“Country-C”、“Country-D”、“1st”、“place” を特徴語 4 0 1 として抽出する（図 2 のステップ 2 1 0）。また、ブロック分割結果 4 3 0 のブロック 1 から、“Sports”、“Championship”、“Cup”、“Country-A”、“broke”、“through”、“primary”、“league”、“first”、“time” が特徴語 4 4 0 として抽出される（図 3 のステップ 3 2 1）。同様に、ブロック 2 から “Country-A”、“played”、“match”、“against”、“first”、“game”、“Country-B”、“Championship”、“ranking”、“highest”、“group”、“though”、“troubled”、“draw” が特徴語 4 4 1 として抽出され、ブロック 3 からは “Country-C”、“game”、“Country-D”、“gained”、“victory”、“offensive”、“strategy”、“passed”、“brilliant”、“group”、“1st”、“place” が特徴語 4 4 2 として抽出され、ブロック 4 からは “final”、“tournament”、“play”、“match”、“against”、“Country-E” が特徴語 4 4 3 として抽出される。

【0 0 3 9】

次に、ブロック別類似度算出プログラム 1 4 1 が実行され、ブロック 1 の特徴語 4 4 0 と種文書の特徴語 4 0 1 から、種文書に対するブロック 1 の類似度を算出する（図 3 のステップ 3 2 2）。本図で示した例では、前記特徴語抽出プログラム 1 5 1 で抽出された種文書の特徴語 4 0 1 とブロック 1 の特徴語 4 4 0 に関して、“Sports”、“Championship”、“Cup”、“Country-A”、“first”、“time” の 6 つの共通の特徴語が存在し、種文書に含まれる特徴語の個数が 1 5 個であることから、前述の式 1 により、“0.40” がブロック 1 の類似度算出結果

450として算出される。

【0040】

同様に、ブロック2～ブロック4についても、それぞれ特徴語抽出プログラム151で抽出された各ブロックの特徴語441～443と種文書の特徴語401から、ブロック別類似度算出プログラム141により種文書に対する各ブロックの類似度“0.33”、“0.33”、“0.00”が類似度算出結果451～453として算出される。

【0041】

次に、上記のブロック1の類似度算出結果450が、あらかじめ設定された種文書適合性判定閾値以上であるか否かを判断し（図3のステップ323）、閾値以上であった場合、ブロック1は種文書に対する適合ブロックと判定し、適合ブロック数を1加算する（図3のステップ324）。本図に示した例では、種文書適合性判定閾値を“0.30”と設定しているためブロック1は適合ブロックと判定され、適合ブロック数と総ブロック数をそれぞれ1加算する（図3のステップ324、325）。

【0042】

同様に、ブロック2～ブロック4についても、図3のステップ323を実行し、ブロック2とブロック3については適合ブロックと判定され、適合ブロック数と総ブロック数が1加算される。またブロック4については非適合ブロックと判定されるため、適合ブロック数は1加算せず総ブロック数のみ1加算される。

【0043】

このように、ブロック1から順に図3のステップ323に示す適合ブロック判定処理を実行した後、適合ブロック数および総ブロック数の算出結果460～463が順に算出され、適合ブロック数および総ブロック数の算出結果463から文書1の適合ブロック数“3”および総ブロック数“4”が算出される。

【0044】

次に、詳細度算出プログラム142が実行され、適合ブロック数および総ブロック数の算出結果463から、前述の式2を用いることにより、文書1の種文書に対する詳細度が“0.75”と算出され（図3のステップ330）、詳細度算出結

果 470 としてワークエリア 160 に格納される（図 3 のステップ 340）。

【0045】

同様に文書 2 に対しても、類似度および詳細度がそれぞれ “1.14”、“0.25” と算出される。

【0046】

磁気ディスク装置 103 に格納されている文書 1 および文書 2 の類似度と詳細度が算出された後、結果出力プログラム 134（図 4 には示していない）が実行され、ワークエリア 160 に格納されている類似度算出結果と詳細度算出結果が、検索結果一覧表示 500（図 5）として出力される。図 5 では、結果出力として、文書 1 および文書 2 に対して文書 ID、類似度、詳細度および見出しが出力されており、文書 1 の類似度および詳細度はそれぞれ “1.06”、“0.75” であり、文書 2 の類似度および詳細度はそれぞれ “1.14”、“0.25” である。

【0047】

ここで、類似度のみでは、文書 1 の類似度 “1.06” と文書 2 の類似度 “1.14” であるから文書 2 の方を有効であると判断してしまうが、文書 1 の詳細度 “0.75” と文書 2 の詳細度 “0.25” から文書 1 の方が文書 2 より種文書に関する内容について全体で適合しているものと判断できる。したがって、出力された詳細度から文書 1 を優先して参照することで効率のいい検索が実現できる。

【0048】

なお、図 5 に示した例では、検索結果一覧表示として文書 ID、類似度、詳細度および見出しを出力するものとしたが、登録処理時に日付など各文書の属性情報も登録しておき、結果出力プログラム 134 でそれらの情報を出力してもよい。また、類似度および詳細度をともに出力するものとしたが、詳細度だけを出力するものとしてもよい。

【0049】

また、図 5 に示した例では、各文書の出力順は類似度の降順で出力するものとしたが、詳細度の降順で出力するものとしてもよいし、これらを図 6 に示すように表示オプションで選択できるようにしてもよい。図 6 に示した例では、表示オプションとして類似度の降順で表示するかあるいは詳細度の降順で表示するかを

選択可能としたインターフェースを備えており、図 6 では詳細度順が選択されていることにより詳細度の高い順に文書 1 と文書 2 が表示されている。

【0 0 5 0】

また、図 5 および図 6 に示した例では、テキスト 1 7 0 として磁気ディスク 1 0 3 に格納されているすべてのテキストに対して結果を表示するものとしたが、図 7 に示すように検索者およびシステム管理者によって予め設定された類似度および詳細度に関する閾値により、検索結果として表示する対象文書を決定してもよい。図 7 に示した例では、類似度および詳細度に関する閾値を設定するインターフェースを備えており、類似度の閾値が“0.00”以上および詳細度の閾値が“0.50”以上と設定されているため、その条件を満たしている文書 1 のみの結果が表示されている。

【0 0 5 1】

また、図 5、図 6 および図 7 に示した例では、類似度および詳細度が検索結果の一覧表示で出力されるものとしたが、図 8 のように指定された文書の全文が表示されるととともに、類似度あるいは詳細度の少なくとも一方が出力されるようにしてもよい。図 8 では、文書 1 の全文を表示するとともに、類似度および詳細度を表示して出力している。また、類似度および詳細度に関してあらかじめ設定された閾値以上の文書に対しては図 8 に示すように類似度、詳細度および全文が出力され、閾値以下の文書に対しては図 5 および図 6 に示すように一覧表示として文書 ID、類似度、詳細度、見出しが出力されるものとしてもよい。

【0 0 5 2】

また、種文書に対する対象文書の類似度算出方法において、類似度算出プログラム 1 3 2 を実行せずに（つまり図 2 のステップ 2 2 2 を実行せずに）、ブロック別類似度算出プログラム 1 4 1 で算出された類似度結果 4 5 0 ～ 4 5 3 を加算することにより（図 3 のステップ 3 2 2）、対象文書の類似度を算出してもよい。

【0 0 5 3】

また、本実施例ではテキスト 1 7 0 のすべてのテキストに対して、類似度算出プログラム 1 3 2（図 2 のステップ 2 2 2）および詳細度算出プログラム 1 3 3

(同ステップ 2 2 3) を実行したが、類似度算出プログラム 1 3 2 で算出された類似度があらかじめ設定された閾値以上のテキストに対して詳細度算出プログラム 1 3 3 を実行してもよい。逆に、詳細度算出プログラム 1 3 3 で算出された詳細度があらかじめ設定された閾値以上のテキストに対して類似度算出プログラム 1 3 2 を実行してもよい。これにより、類似度あるいは詳細度の算出対象となるテキスト数を削減することができ、高速に検索を行うことができる。

【 0 0 5 4 】

また、本実施例では、予め蓄積された文書に対して検索条件との関連性を判定する文書検索システムとして説明したが、特開平 2 0 0 0 - 3 3 9 3 4 6 号公報に記載されている類似文書検索配送システムにおける適合度算出プログラムを、本発明における詳細度算出制御プログラムに置き換えてもよい。

【 0 0 5 5 】

このように本発明による詳細度は、予め蓄積された文書に対して検索条件との関連性を判定する文書検索システムだけでなく、1 件の対象文書に対して配信条件との関連性を判定する文書配信システムにも適用できる。

【 0 0 5 6 】

以上説明したように、本発明の第一の実施例によれば、種文書に関する内容について、対象文書の全体で類似しているのか、あるいは対象文書の一部で類似しているのかを判断できるため、有効な文書を効率よく検索できるようになる。

【 0 0 5 7 】

次に、本発明の第二の実施例について説明する。第二の実施例では、検索条件として種文書と全文検索条件の両方が指定された場合における詳細度を算出する。

【 0 0 5 8 】

本実施例の文書検索システムは、図 1 に示した第一の実施例のシステムとほぼ同様の構成であるが、検索制御プログラム 1 1 2 と詳細度算出プログラム 1 3 3 が異なり、図 9 に示すように、検索制御プログラム 1 1 2 c には全文検索条件解析プログラム 1 3 0 a が加わるとともに、詳細度算出プログラム 1 3 3 0 にはブロック別全文検索条件適合度算出プログラム 1 4 1 a が加わる。

【0059】

以下、第一の実施例と異なる検索制御プログラム 112c の処理手順について図 10 を用いて説明する。ここで第一の実施例（図 2）と異なるのは、種文書解析プログラム 130 が実行された後に全文検索条件解析プログラム 130a が実行されること、及び類似度算出プログラム 132 が実行された後に詳細度算出制御プログラム 1330 が実行されることである。

【0060】

検索制御プログラム 112c は、まず種文書解析プログラム 130 を起動し、検索条件で指定された種文書を読み込み、ワークエリア 160 に格納する（ステップ 200）。次に、全文検索条件解析プログラム 130a を起動し、検索条件で指定された全文検索条件を読み込む。この全文検索条件に含まれる AND、OR、NOT の論理演算子を識別することによりその構造を解析し、和積標準形で表された論理演算式（以下、解析済論理演算式と呼ぶ）をワークエリア 160 に格納する（ステップ 200a）。次に、特徴語抽出プログラム 151 を起動し、前記種文書解析プログラム 130 によりワークエリア 160 に格納された種文書から特徴語を抽出し、ワークエリア 160 に格納する（ステップ 210）。

【0061】

次に、テキスト 170 に含まれるすべてのテキストに対して、ステップ 221 ～ステップ 223 を繰り返し実行する（ステップ 220）。まず、テキスト読込プログラム 131 を起動し、磁気ディスク装置 103 に格納されているテキスト 170 からテキストを 1 つ読み込む（ステップ 221）。次に、類似度算出プログラム 132 を起動し、前記テキスト読込プログラム 131 により読み込まれたテキストに対し、種文書に対するテキストの類似度を算出し、ワークエリア 160 に格納する（ステップ 222）。詳細度算出制御プログラム 1330 を起動し、検索条件に対する前記テキスト読込プログラム 131 により読み込まれたテキストの詳細度を算出し、ワークエリア 160 に格納する（ステップ 223c）。

【0062】

そして、結果出力プログラム 134 を起動し、前記類似度算出プログラム 132 により算出された類似度と前記詳細度算出制御プログラム 1330 により算出

された詳細度を各テキストに対して出力する（ステップ230）。

【0063】

次に、詳細度算出制御プログラム1330の処理手順について図11を用いて説明する。ここで第一の実施例（図3）と異なるのは、ブロック別類似度算出プログラム141が実行された後にブロック別全文検索条件適合度算出プログラム141aが実行されることと、図3に示す適合性判定ステップ323において、種文書適合性判定閾値のみを適合ブロック判定基準に用いるのではなく、ブロック別全文検索条件適合度算出プログラム141aによって算出された全文検索条件適合度に関する閾値（以下、全文検索条件適合性判定閾値と呼ぶ）も適合ブロック判定基準に用いることである。

【0064】

まず、テキストの適合ブロック数とテキストに含まれる総ブロック数の初期値とともに0に設定する（ステップ300）。ブロック分割プログラム140を起動し、ステップ221（図10）において読み込まれたテキストをブロックに分割する（ステップ310）。

【0065】

次に、ステップ310で分割された各ブロックに対して、それぞれステップ321～325を繰り返し実行する（ステップ320）。まず、特徴語抽出プログラム151を起動し、各ブロックから特徴語を抽出する（ステップ321）。次に、ブロック別類似度算出プログラム141を起動し、特徴語抽出プログラム151により抽出された種文書の特徴語と前記ステップ321で抽出された各ブロックの特徴語から、種文書に対するブロックの類似度を式1を用いて算出する（ステップ322）。

【0066】

【式 1】

式1

$$\text{類似度} = \frac{\text{種文書とブロックの共通の特徴語の数}}{\text{種文書の特徴語の数}}$$

次に、ブロック別全文検索条件適合度算出プログラム 141a を起動し、全文検索条件解析プログラム 130a により読み込まれた解析済論理演算式から、全文検索条件に対するブロックの適合度（以下、全文検索条件適合度と呼ぶ）を算出する（ステップ 322a）。

【0067】

次に、前記ブロック別類似度算出プログラム 141 により算出された各ブロックの類似度を、種文書適合性判定閾値と比較するとともに、ステップ 322a で算出されたブロックの全文検索条件適合度を、全文検索条件適合性判定閾値と比較する（ステップ 323c）。この比較の結果、あるブロックの類似度が種文書適合性判定閾値以上であり、かつそのブロックの全文検索条件適合度が全文検索条件適合性判定閾値以上の場合、そのブロックを検索条件に対する適合ブロックと判定し、適合ブロック数を 1 加算するとともに（ステップ 324）、総ブロック数を 1 加算する（ステップ 325）。ステップ 323c において適合度または類似度のどちらかが閾値以下であった場合は、適合ブロック数は 1 加算されず、総ブロック数のみが 1 加算される（ステップ 325）。

【0068】

次に、詳細度算出プログラム 142 を起動し、前記ステップ 324 およびステップ 325 で計数された適合ブロック数と総ブロック数から、式 2 を用いて種文書に対する該テキストの詳細度を算出する（ステップ 330）。

【0069】

【式 2】

式 2

$$\text{詳細度} = \frac{\text{適合ブロック数}}{\text{総ブロック数}}$$

最後に、前記ステップ 3 3 0 で算出された種文書に対する該テキストの詳細度をワークエリア 1 6 0 に格納する（ステップ 3 4 0）。

【0 0 7 0】

次に、詳細度算出制御プログラム 1 3 3 0 により起動されるブロック別全文検索条件適合度算出プログラム 1 4 1 a の処理手順について説明する。まず、全文検索条件解析プログラム 1 3 0 a により和積標準形でワークエリア 1 6 0 に読み込まれた解析済論理演算式に対し、AND 演算子を境界として分割される単語や論理演算式（以下、部分論理演算式）を抽出する。次に、特徴語抽出プログラム 1 5 1 により抽出された処理対象となるブロックの特徴語が、抽出された各部分論理式の条件と適合するかどうかを判定する。

【0 0 7 1】

この結果、処理対象のブロックが満たす部分論理演算式の数（以下、適合部分論理式数と呼ぶ）と、解析済論理演算式に含まれる部分論理演算式（以下、総部分論理演算式数と呼ぶ）を計数し、式 3 より全文検索条件に対するブロックの全文検索条件適合度を算出する。

【0 0 7 2】

【式 3】

式3

$$\text{全文検索条件適合度} = \frac{\text{適合部分論理演算式数}}{\text{総部分論理演算式数}}$$

なお、ステップ 3 2 2 a における、ブロック別全文検索条件適合度算出プログラム 1 4 1 a によるブロックの全文検索条件適合度の算出には、指定された全文検索条件に含まれる部分論理演算式の総数に対し、該ブロックの特徴語により満たされている部分論理式の数の割合を算出したが、特開平 1 1 - 1 5 4 1 6 4 号公報や特開 2 0 0 1 - 8 4 2 5 5 号公報に開示されている方法を用いてもよい。

【0 0 7 3】

以下、本実施例の検索処理におけるブロックの適合性判定について、具体的な処理の流れを図 1 2 を用いて説明する。

【0 0 7 4】

本図に示した例は、文書 1 「In The Sports Championship Cup, Country-A broke through the primary league for the first time. Country-A played a match against Country-B of the Championship ranking highest in H group at the first game, and though troubled, and was a draw. Then, both the Country-C game and the Country-D game gained a victory with offensive strategy, and passed the brilliant H group by the 1st place. A final tournament is due to play a match against Country-E.」が磁気ディスク装置 1 0 3 に格納された文書検索システムにおいて、種文書として「The Sports Championship Cup held for the first time in Country-A , and Country-A passed H group including Country-B, Country-C, and Country-D by the 1st place.」、全文検索条件として「“Country-A” and “Country-B” and (“Championship” or “tournament”)」が入力された場合の例を示している。なお、本図は、種文書解析プログラム 1 3 0 により検索条件として入力された種文書が文書 4 0 0 と

して読み込まれ、全文検索条件解析プログラム 130a により検索条件として入力された全文検索条件が解析済論理演算式 4000 として読み込まれ、テキスト読込プログラム 131 により文書 1 がテキスト 410 として読み込まれた状態である。

【0075】

まず、特徴語抽出プログラム 151 が実行され、前記種文書解析プログラム 130 により読み込まれた種文書 400 から、“Sports”、“Championship”、“Cup”、“held”、“first”、“time”、“Country-A”、“passed”、“group”、“including”、“Country-B”、“Country-C”、“Country-D”、“1st”、“place” を特徴語 401 として抽出する（図 10 のステップ 210）。次に、ブロック分割プログラム 140 が実行され、テキスト 410 をブロック単位へ分割する（図 11 のステップ 310）。本図に示した例では、テキスト 410 を“.”（ピリオド）を区切り文字としてブロック単位に分割しており、この分割結果から「In The Sports Championship Cup, Country-A broke through the primary league for the first time.」がブロック 1 の抽出結果 4300 として出力されている。

【0076】

次に、特徴語抽出プログラム 151 が実行され、ブロック分割プログラム 140 で文書 1 より分割されたブロック 1 から、“Sports”、“Championship”、“Cup”、“Country-A”、“broke”、“through”、“primary”、“league”、“first”、“time” を特徴語 440 として抽出する（図 11 のステップ 321）。次に、ブロック別類似度算出プログラム 141 が実行され、ブロック 1 の特徴語 440 と種文書の特徴語 401 から、種文書に対するブロック 1 の類似度を算出する（図 11 のステップ 322）。本図で示した例では、特徴語抽出プログラム 151 で抽出された種文書の特徴語 401 とブロック 1 の特徴語 440 の間で、“Sports”、“Championship”、“Cup”、“Country-A”、“first”、“time” の 6 つの共通の特徴語が存在し、種文書に含まれる特徴語の個数が 15 個であることから、前述した式 1 より、“0.40” がブロック 1 の類似度算出結果 450 として算出される。

【0077】

次に、ブロック別全文検索条件適合度算出プログラム141aが実行され、全文検索条件に対するブロック1の全文検索条件適合度を算出する（図11の322a）。本図で示した例では、ブロック1の特徴語440には“Country-A”および“Championship”が含まれており、解析済論理演算式4000「“Country-A” and “Country-B” and (“Championship” or “tournament”)」の部分論理演算式「“Country-A”」、「“Championship” or “tournament”」を満たしている。すなわち、解析済論理演算式4000に含まれる3つの部分論理演算式のうち、2つが満たされていることから、“0.67”がブロック1の全文検索条件適合度算出結果4500として算出される。

【0078】

そして、ブロック1の類似度が種文書適合性判定閾値以上であり、かつブロック1の全文検索条件適合度が全文検索条件適合性閾値以上であるかどうかを判定する（図11のステップ323c）。判定の結果、両方の値が閾値以上である場合は、ブロック1は検索条件に対して適合ブロックと判定される。本図に示した例では、種文書適合性閾値および全文検索条件適合性閾値をそれぞれ“0.30”としており、ブロック1の類似度“0.40”、および詳細度“0.67”はそれぞれこの条件を満たしているため適合ブロックと判定される。

【0079】

次に、図12に示した、ブロック別全文検索条件適合度算出プログラム141aが行うブロック別全文検索条件適合度算出処理（図11のステップ322a）の詳細について、図13を用いて説明する。

【0080】

本図に示した例では、全文検索条件解析プログラム130aによって読み込まれた解析済論理式4000「“Country-A” and “Country-B” and (“Championship” or “tournament”)」に対し、図12に示したブロック1の特徴語440からブロック1の全文検索適合度を算出する場合の処理の流れを示している。

【0081】

まず、解析済論理演算式4000から部分論理演算式4501を抽出する（ス

テップ 3 2 2 1)。ここでは、和積標準形で読み込まれた解析済論理演算式が AND 演算子を境界として分割され、その分割された単語や論理演算式を、部分論理式として抽出する。本図に示した例では、AND 演算子を境界として、解析済論理演算式 4 0 0 0 から「“Country-A”」、「“Country-B”」、「“Championship” or “tournament”」が抽出される。

【0082】

次に、ブロック 1 の特徴語 4 4 0 と前記部分論理演算式抽出ステップ 3 2 2 1 によって抽出された部分論理演算式 4 5 0 1 から、各部分論理演算式に対するブロックの適合判定を行う（ステップ 3 2 2 1）。そして、判定結果 4 5 0 2 を出力する。本図に示した例では、ブロック 1 の特徴語が“Country-A”、“Championship”を含むことから、ブロック 1 を満たす部分論理演算式 4 5 0 1 は「“Country-A”」、「“Championship” or “tournament”」と判定される。

【0083】

次に、解析済論理演算式 4 0 0 0 に対するブロック 1 の全文検索条件適合度 4 5 0 0 を算出する（ステップ 3 2 2 3）。本図に示した例では、前記部分論理演算式適合判定ステップ 3 2 2 2 によるブロックの適合判定結果 4 5 0 2 から、部分論理式数“3”が計数されると共に、ブロック 1 が満たす部分論理式数“2”と計数される。この結果、式 3 より“0.67”が全文検索条件適合度 4 5 0 0 として算出される。

【0084】

以上説明したように本発明の第二の実施形態によれば、種文書の内容に対する類似性と全文検索条件に対する適合性の両方を用いて詳細度の算出を行うことにより、検索者の検索目的に応じた、より精度の高い検索条件に関する文書の詳細度を算出することができる。

【0085】

なお本実施例では、検索条件として種文書と全文検索条件の両方を指定する構成を採用したが、全文検索条件のみが指定される場合でもよい。その場合、図 9 に示した種文書解析プログラム 1 3 0 とブロック別類似度算出プログラム 1 4 1 がなくなるとともに、図 1 1 に示したステップ 3 2 3 c の適合ブロックの判定処

理に関する判定基準が全文検索条件適合度のみとなる。また、図 1 0 に示したステップ 2 2 2 における類似度算出処理は、全文検索条件に関するテキストの類似度として、拡張ブーリアンに基づいた方法や、特開平 1 1 - 1 5 4 1 6 4 号公報に基づいた方法で算出される。

【 0 0 8 6 】

次に、第三の実施例について説明する。第三の実施例では、文書ファイルの登録時にブロックごとに抽出された特徴語を、あらかじめブロック別特徴語ファイルとして格納しておき、詳細度の算出時には、そのブロック別特徴語ファイルを読み込むことで詳細度を算出する。

【 0 0 8 7 】

本実施例の文書検索システムは、図 1 に示した第一の実施例のシステムとほぼ同様の構成を取るが、図 1 4 に示すように磁気ディスク装置 1 0 3 にブロック別特徴語ファイル 1 7 1 が追加されるとともに、登録制御プログラム 1 1 1 と詳細度算出制御プログラム 1 3 3 の構成が異なり、登録制御プログラム 1 1 1 c にはブロック分割プログラム 1 4 0 とブロック別特徴語登録プログラム 1 2 0 0 が加わるとともに、詳細度算出制御プログラム 1 3 3 1 にはブロック分割プログラム 1 4 0 の代りに特徴語読込プログラム 1 4 0 0 が加わる。

【 0 0 8 8 】

以下、第一の実施例とは異なる登録制御プログラム 1 1 1 c の処理手順を図 1 5 を用いて説明する。ここで、第一の実施例と異なるのは、テキスト登録プログラム 1 2 1 が実行された後に、ブロック別特徴語ファイル 1 7 1 を作成するために、ブロック分割プログラム 1 4 0、特徴語抽出プログラム 1 5 1 およびブロック別特徴語登録プログラム 1 2 0 0 が実行されることである。

【 0 0 8 9 】

登録制御プログラム 1 1 1 c では、まず文書ファイル取得プログラム 1 2 0 を起動し、FDD 1 0 4 を介してフレキシブルディスク 1 0 8 に格納されている文書ファイルをワークエリア 1 6 0 に読み込む（ステップ 7 0 0）。次に、テキスト登録プログラム 1 2 1 を起動して、ステップ 7 0 0 で読み込まれた文書ファイルからテキストを抽出し、ワークエリア 1 6 0 に格納するとともにテキスト 1 7 0

として磁気ディスク装置 103 に格納する（ステップ 710）。次に、ブロック分割プログラム 140 を起動し、ステップ 710 でワークエリア 160 に格納されたテキストをブロック単位に分割する（ステップ 720）。

【0090】

次に、ステップ 720 で分割された各ブロックに対して、それぞれステップ 731～ステップ 732 を繰り返し行う（ステップ 730）。まず、特徴語抽出プログラム 151 を起動し、各ブロックの特徴語を抽出する（ステップ 731）。次に、ブロック別特徴語ファイル作成プログラム 1200 を起動し、ステップ 731 により各ブロックから抽出された特徴語を、ブロック別特徴語ファイル 171 に登録する（ステップ 732）。

【0091】

以下、第一の実施例と異なる詳細度算出制御プログラム 1331 の処理手順を図 16 を用いて説明する。第一の実施例における詳細度算出制御プログラム 133 の処理手順（図 3）と異なるのは、ステップ 310 がなくなるとともに、ステップ 321 の代りにステップ 321a が加わることである。

【0092】

まず、詳細度算出制御プログラム 1331 は、まず適合ブロック数と総ブロック数の初期値をともに 0 と設定する（ステップ 300）。次に、1 つのテキストに含まれるすべてブロックに対して、それぞれステップ 321a～ステップ 325 を繰り返し実行する（ステップ 320）。

【0093】

まず、特徴語読込プログラム 1400 を起動し、ブロック別特徴語ファイル 171 から 1 ブロック分の特徴語を読み込む（ステップ 321a）。次に、ブロック別類似度算出プログラム 141 を起動し、上述した式 1 より種文書に対するブロックの類似度を算出する（ステップ 322）。次に、ステップ 322 で算出されたブロックの類似度を種文書適合性判定閾値と比較する（ステップ 323）。この結果、ブロックの類似度が種文書適合性判定閾値以上であった場合、そのブロックは適合ブロックと判定され、適合ブロック数を 1 加算するとともに（ステップ 324）、総ブロック数を 1 加算する（ステップ 325）。ステップ 323

において閾値以下であった場合は、適合ブロック数は 1 加算されず、総ブロック数のみが 1 加算される（ステップ 3 2 5）。

【 0 0 9 4 】

次に、詳細度算出プログラム 1 4 2 を起動し、ステップ 3 2 4 およびステップ 3 2 5 で計数された適合ブロック数と総ブロック数から、式 2 を用いて種文書に対するそのテキストの詳細度を算出する（ステップ 3 3 0）。次に、ステップ 3 3 0 で算出された種文書に対するそのテキストの詳細度をワークエリア 1 6 0 に格納する（ステップ 3 4 0）。

【 0 0 9 5 】

次に、文書の登録処理におけるブロック別の特徴語をディスク装置 1 0 3 のブロック別特徴語ファイル 1 7 1 に登録する処理の流れについて、図 1 7 を用いて説明する。本図に示した例では、文書 1 「In The Sports Championship Cup, Country-A broke through the primary league for the first time. Country-A played a match against Country-B of the Championship ranking highest in H group at the first game, and though troubled, and was a draw. Then, both the Country-C game and the Country-D game gained a victory with offensive strategy, and passed the brilliant H group by the 1st place. A final tournament is due to play a match against Country-E.」 および文書 2 「Country-A is still in the state of economic depression. If there are bright news that induce an economic big effect, can Country-A escape from economic depression? The Sports Championship Cup was held for the first time in Country-A, and Country-A passed H group including Country-B, Country-C, and Country-D by the 1st place on the other day. However, it was not able to become an explosive to economic recovery and an economic big effect could not be acquired.」 が、テキスト読込プログラム 1 3 1 により、それぞれテキスト 4 1 0 およびテキスト 9 0 0 として読み込まれた状態から、文書 1 および文書 2 の各ブロックの特徴語をブロック別特徴語ファイル 1 7 1 に登録する処理の流れを説明している。

【 0 0 9 6 】

まず、ブロック分割プログラム 140 が実行され、テキスト読込プログラム 131 により読み込まれたテキスト 410 をブロック単位に分割する。本図に示した例では、“.”（ピリオド）を区切り文字としてテキスト 410 をブロック単位に分割しており、この結果としてブロック分割結果 430 が出力される。図 17 に示したブロック分割結果 430 は、ブロック 1 「In The Sports Championship Cup, Country-A broke through the primary league for the first time.」、ブロック 2 「Country-A played a match against the first game and Country-B of the Championship ranking highest in H group, and though troubled, and was a draw.」、ブロック 3 「Then, both the Country-C game and the Country-D game gained a victory with offensive strategy, and passed the brilliant H group by the 1st place.」およびブロック 4 「A final tournament is due to play a match against Country-E.」が格納されていることを表している。

【0097】

次に、特徴語抽出プログラム 151 が実行され、ブロック分割結果 430 のブロック 1 から、特徴語 440 として “Sports”、“Championship”、“Cup”、“Country-A”、“broke”、“through”、“primary”、“league”、“first”、“time” を抽出する。そして、ブロック別特徴語登録プログラム 1200 が実行され、前記特徴語抽出プログラム 151 により抽出されたブロック 1 の特徴語 440 は、文書 1 のブロック 1 の特徴語として、ブロック別特徴語ファイル 171 に登録される。また、合わせて文書 ID “1” およびブロック ID “1” もブロック別特徴語ファイル 171 に登録される。

【0098】

同様にブロック 2～ブロック 4 についても特徴語抽出プログラム 151 により特徴語 441～443 が抽出され、各ブロックにおいて抽出された特徴語がそれぞれ文書 1 の各ブロックの特徴語としてブロック別特徴語ファイル 171 に登録される。

【0099】

同様に文書 2 についても、テキスト読込プログラム 131 で読み込まれたテキ

スト 9 0 0 に対し、ブロック分割プログラム 1 4 0 によりブロック分割結果 9 0 1 が出力され、特徴語抽出プログラム 1 5 1 により各ブロックから特徴語 9 4 0 ~ 9 4 3 が抽出され、抽出された特徴語が、ブロック別特徴語登録プログラム 1 2 0 0 によりそれぞれ文書 2 の各ブロックの特徴語としてブロック別特徴語ファイル 1 7 1 に登録される。

【0 1 0 0】

なお、本図のブロック別特徴語ファイル 1 7 1 に格納されている文書 ID “1” および “2” は、それぞれ文書 1 および文書 2 に対応している。

【0 1 0 1】

以上説明したように、本発明の第三の実施例によれば、ブロック別特徴語ファイル 1 7 1 を文書登録時にあらかじめ作成しておくことにより、検索の度にテキストのブロック分割処理およびブロックの特徴語抽出処理を実行する必要がないため、検索時には大量のテキストに対しても高速に詳細度の算出を行うことができる。

【0 1 0 2】

なお、本実施例においては、テキスト読込プログラム 1 3 1 を起動してテキスト 1 7 0 を読み込み、類似度を算出する構成としたが、テキスト読込プログラム 1 3 1 を呼び出さず、検索制御プログラム 1 1 2 が特徴語読込プログラム 1 4 0 を呼び出し、ブロック別特徴語ファイル 1 7 1 を読み込んだ値を用いて類似度を算出してもよい。これにより、テキストを読み込まなくてもよくなるため、メモリの使用量を軽減することができる。

【0 1 0 3】

【発明の効果】

以上説明したように本発明によれば、種文書に関する対象文書の類似度だけでなく、対象文書全体に対して種文書の内容が占める割合を表す詳細度が出力されるようになる。これにより、種文書に関する内容について、対象文書の全体で類似しているのか、あるいは対象文書の一部で類似しているのかを容易に判断できるため、文書を効率よく検索できる。

【図面の簡単な説明】

【図 1】本発明の第一の実施例における類似文書検索システムの全体構成を示す図である。

【図 2】本発明の第一の実施例における検索制御プログラム 112 の処理を示す PAD 図である。

【図 3】本発明の第一の実施例における詳細度算出制御プログラム 133 の処理を説明する PAD 図である。

【図 4】本発明の第一の実施例における検索制御プログラム 112 の具体的な処理の流れを説明する図である。

【図 5】本発明の第一の実施例における検索結果一覧画面を示す図である。

【図 6】本発明の第一の実施例における検索結果一覧画面を示す図である。

【図 7】本発明の第一の実施例における結果出力プログラム 134 の出力対象文書として、類似度および詳細度の閾値を設定する検索結果一覧画面を示す図である。

【図 8】本発明の第一の実施例における、対象文書の全文を表示する画面を示す図である。

【図 9】本発明の第二の実施例における類似文書検索システムの全体構成を示す図である。

【図 10】本発明の第二の実施例における検索制御プログラム 112c の処理を説明する PAD 図である。

【図 11】本発明の第二の実施例における詳細度算出制御プログラム 1330 の処理を説明する PAD 図である。

【図 12】本発明の第二の実施例の検索制御プログラム 112c における適合ブロック判定処理の具体的な流れを説明する図である。

【図 13】本発明の第二の実施例の全文検索条件適合度算出プログラム 141a の具体的な処理の流れを説明する図である。

【図 14】本発明の第三の実施例における類似文書検索システムの全体構成を示す図である。

【図 15】本発明の第三の実施例における登録制御プログラム 111 の処理を説明する PAD 図である。

【図 1 6】本発明の第三の実施例における詳細度算出制御プログラム 1 3 3 1 の処理を説明する P A D 図である。

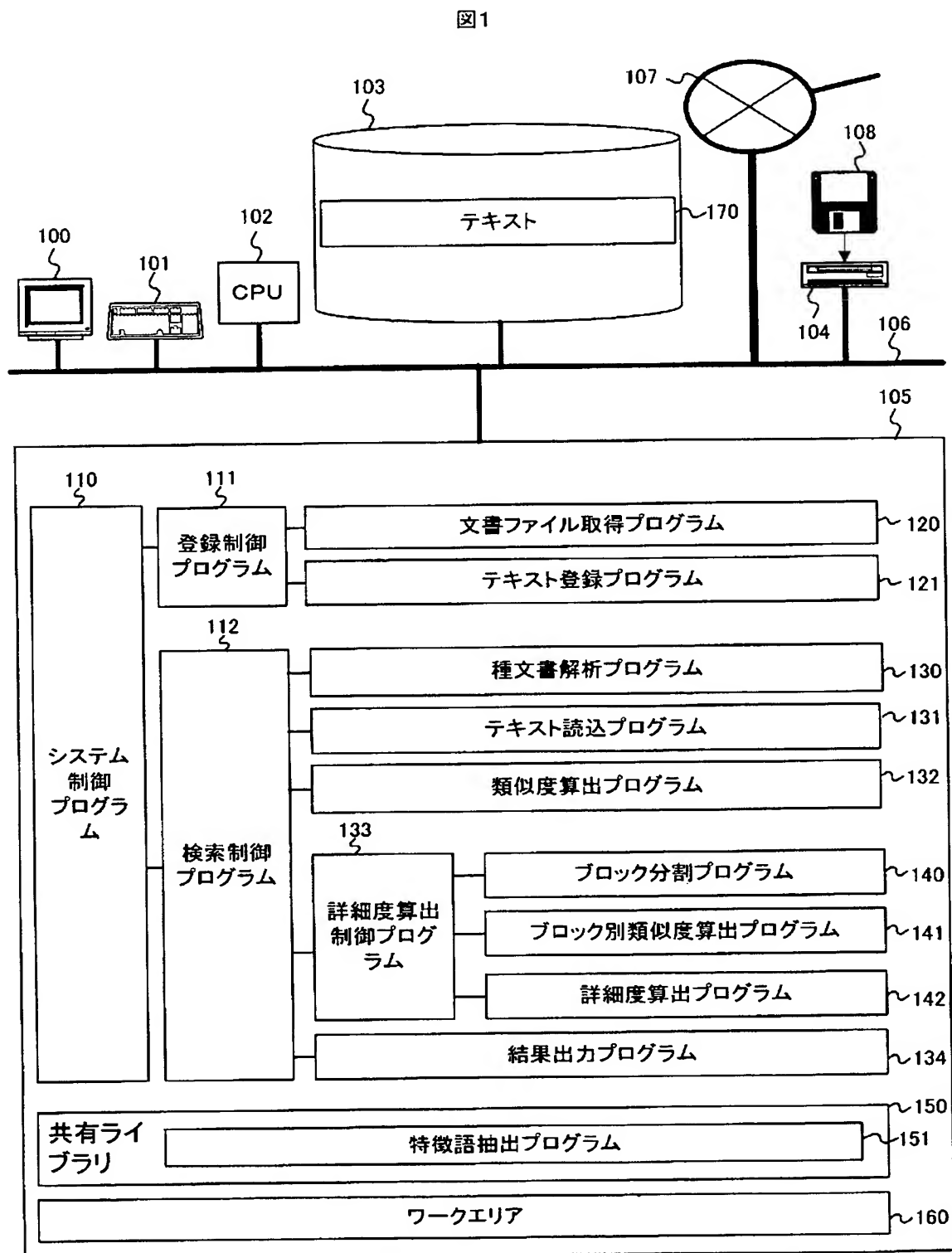
【図 1 7】本発明の第三の実施例における登録制御プログラム 1 1 1 の具体的な処理の流れを説明する図である。

【符号の説明】

1 0 0…ディスプレイ、1 0 1…キーボード、1 0 2…中央演算処理装置（C P U）、1 0 3…磁気ディスク装置、1 0 4…フレキシブルディスクドライブ（F D D）、1 0 5…主メモリ、1 1 0…システム制御プログラム、1 1 1…登録制御プログラム、1 1 2…検索制御プログラム、1 2 0…文書ファイル取得ファイル、1 2 1…テキスト登録プログラム、1 3 0…種文書解析プログラム、1 3 1…テキスト読込プログラム、1 3 2…類似度算出プログラム、1 3 3…詳細度算出制御プログラム、1 3 4…結果出力プログラム、1 4 0…ブロック分割プログラム、1 4 1…ブロック別類似度算出プログラム、1 4 2…詳細度算出プログラム、1 5 0…共有ライブラリ、1 5 1…特徴語抽出プログラム、

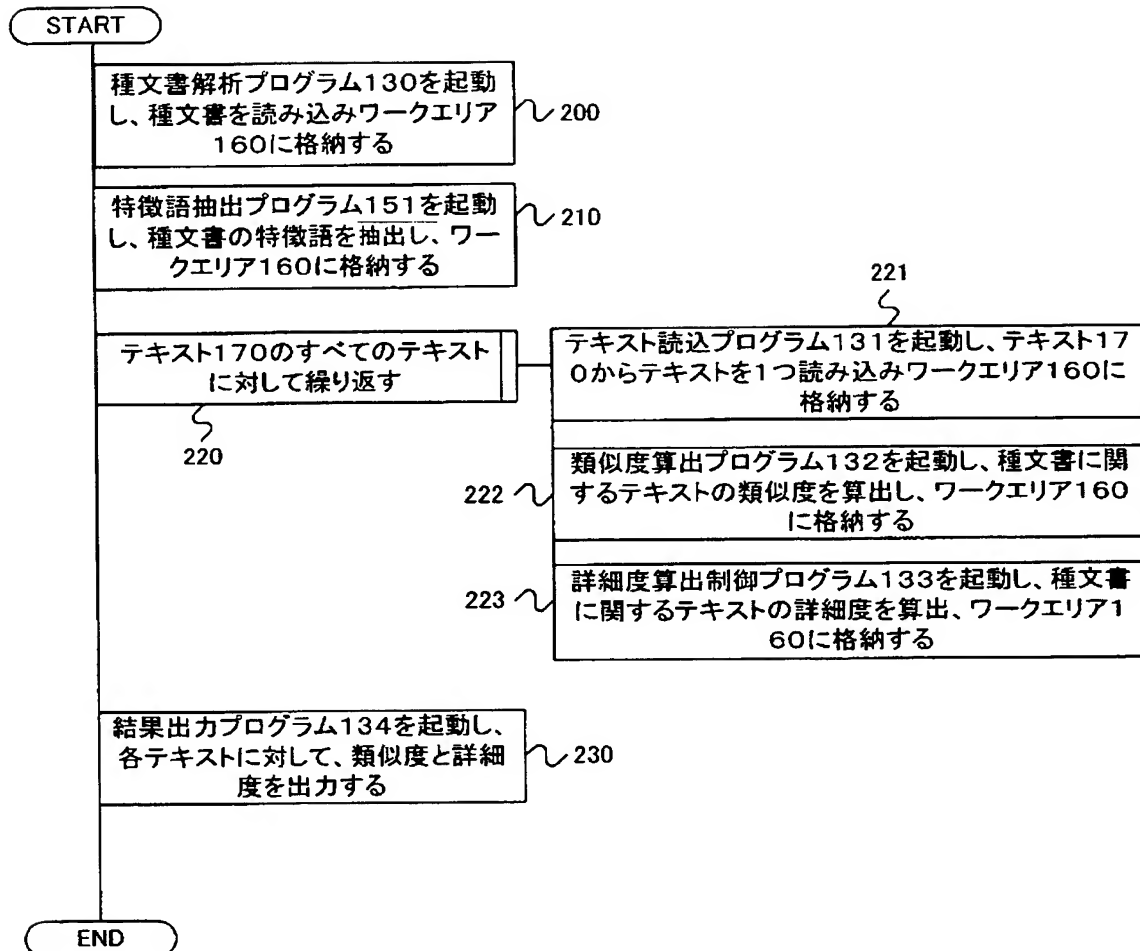
【書類名】 図面

【図 1】



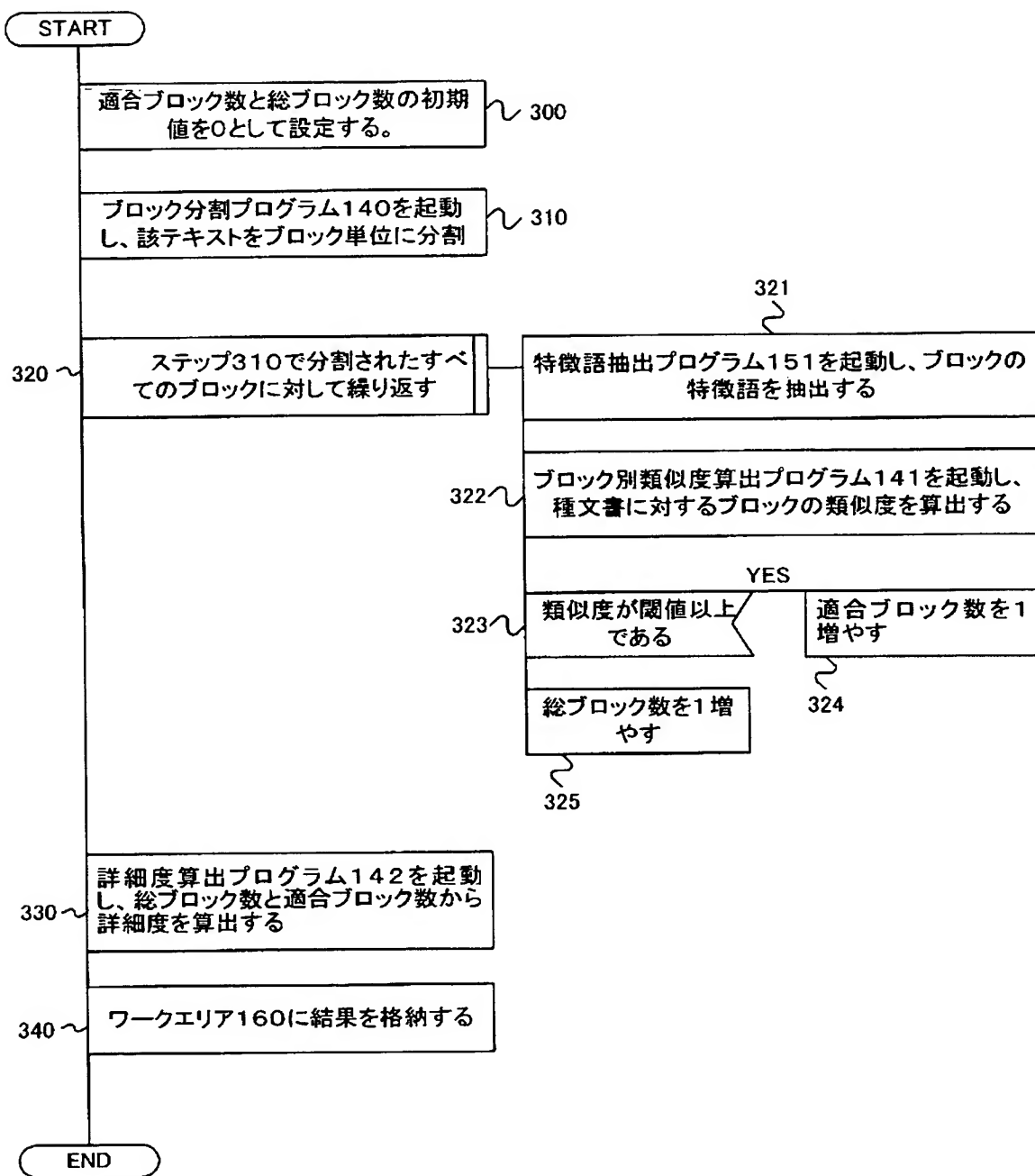
【図 2】

図2



【図 3】

図3



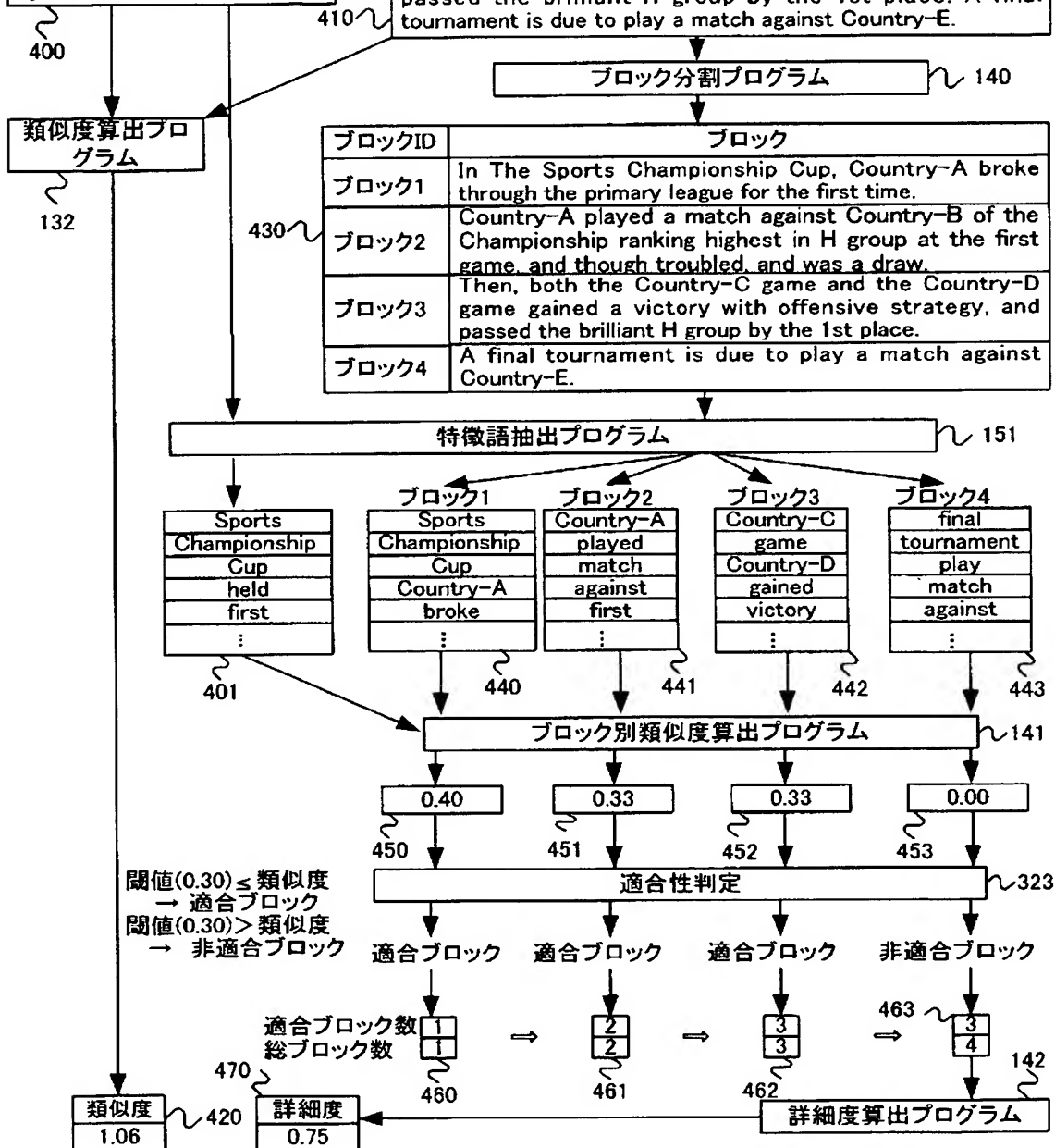
【図 4】

種文書

The Sports Championship Cup was held for the first time in Country A, and Country-A passed H group including Country-B, Country-C, and Country-D by the 1st place.

In The Sports Championship Cup, Country-A broke through the primary league for the first time. Country-A played a match against Country-B of the Championship ranking highest in H group at the first game, and though troubled, and was a draw. Then, both the Country-C game and the Country-D game gained a victory with offensive strategy, and passed the brilliant H group by the 1st place. A final tournament is due to play a match against Country-E.

図4



【図 5】

図5

検索結果一覧			
文書ID	類似度	詳細度	見出し
2	1.14	0.25	Country-A is still in the state of economic ...
1	1.06	0.75	In The Sports Championship Cup, Country-A ...

【図 6】

図6

検索結果一覧			
表示順 <input type="radio"/> 類似度順 <input checked="" type="radio"/> 詳細度順			
文書ID	詳細度	類似度	見出し
1	0.75	1.06	In The Sports Championship Cup, Country-A ...
2	0.25	1.14	Country-A is still in the state of economic ...

【図 7】

図7

検索結果一覧			
表示文書条件			
類似度 <input type="text" value="0"/> 以上 詳細度 <input type="text" value="0.5"/> 以上			
文書ID	類似度	詳細度	見出し
1	1.06	0.75	In The Sports Championship Cup, Country-A ...

【図 8】

図8

文書1 本文表示

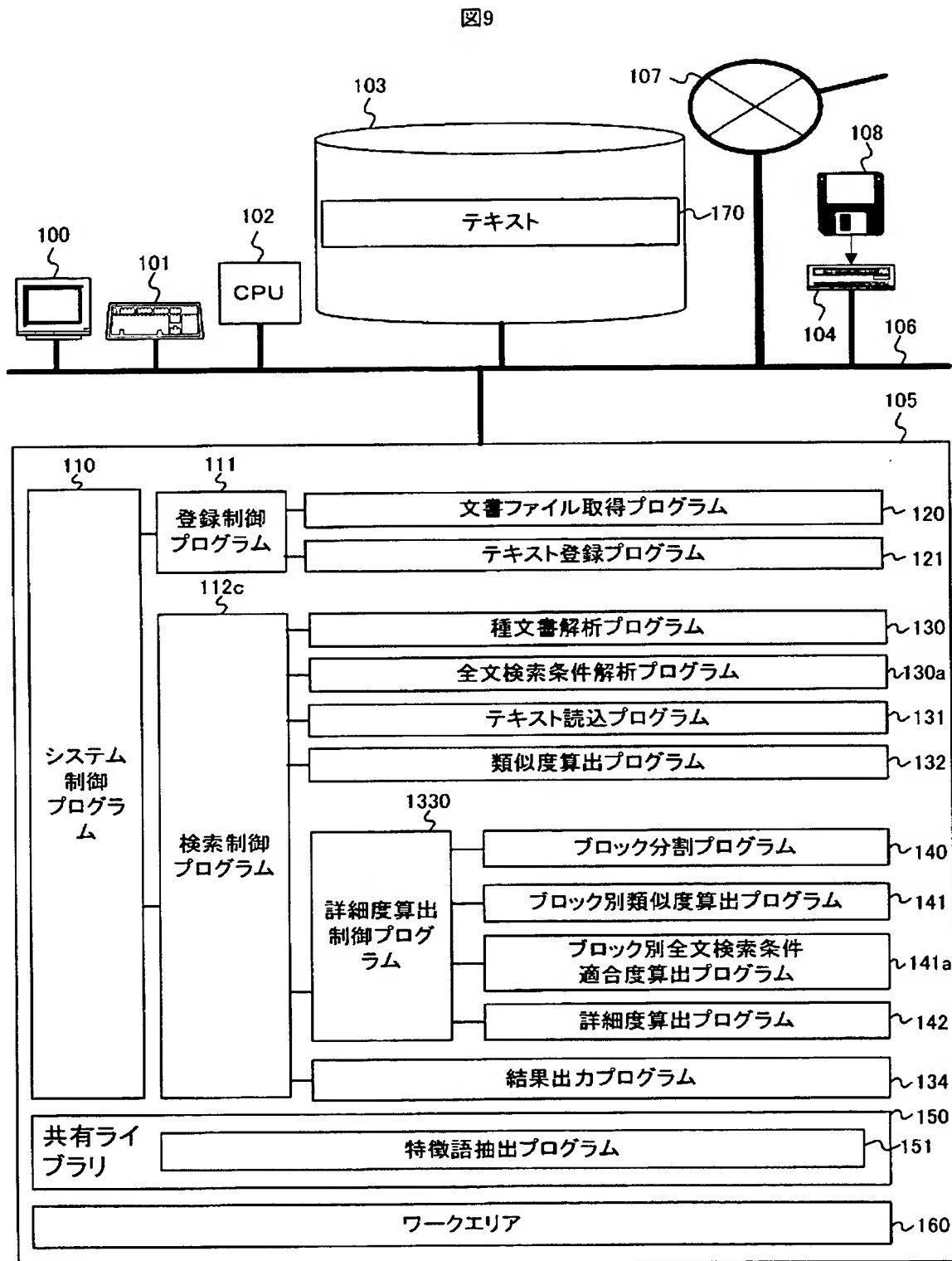
類似度 1.06

詳細度 0.75

In The Sports Championship Cup, Country-A broke through the primary league for the first time. Country-A played a match against the first game and Country-B of the world ranking highest in H group, and though troubled, and was a draw. Then, both the Country-C game and the Country-D game gained a victory with offensive strategy, and passed the brilliant H group by the 1st place. A final tournament is due to play a match against Country-E.

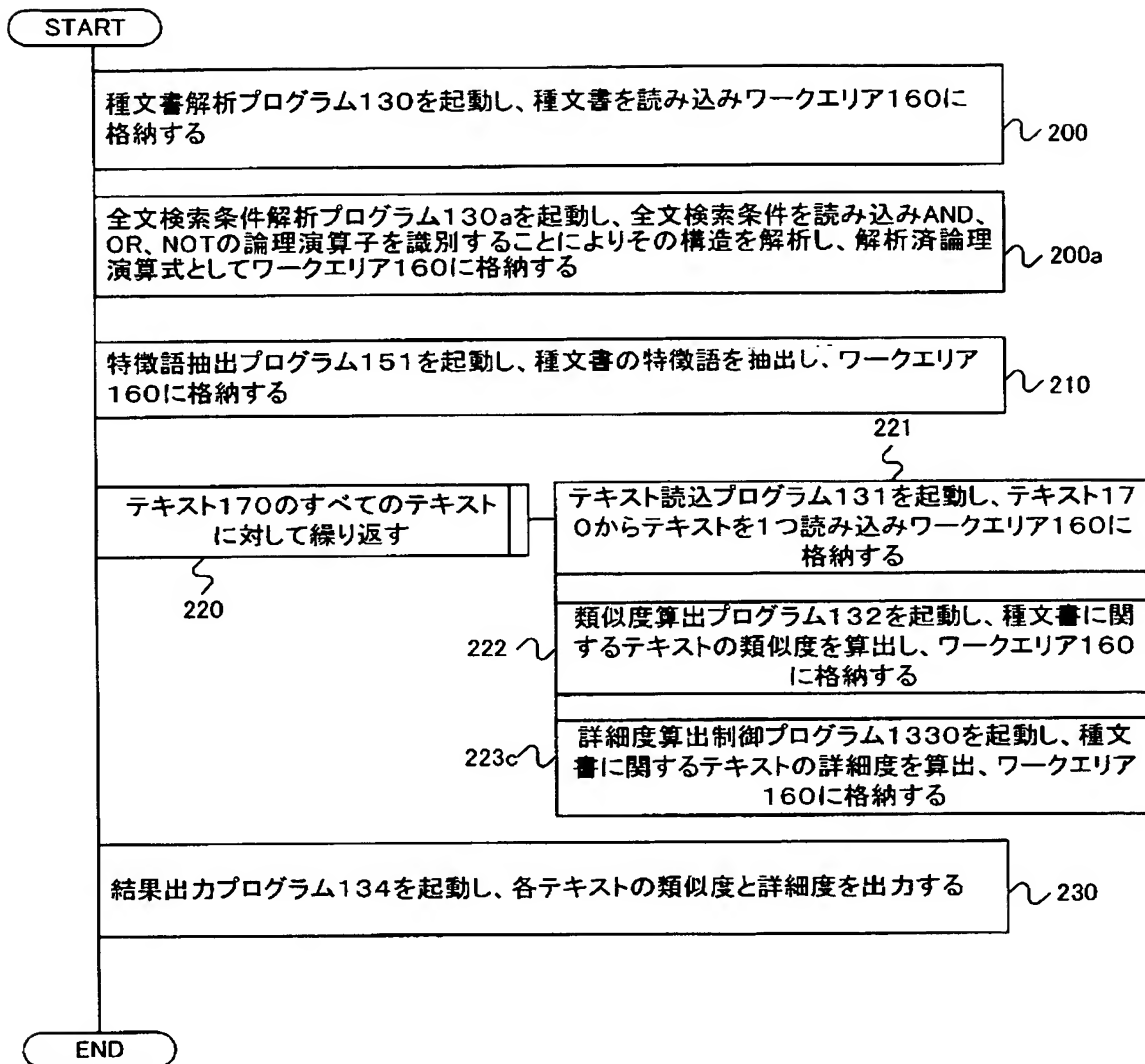
503

【図 9】



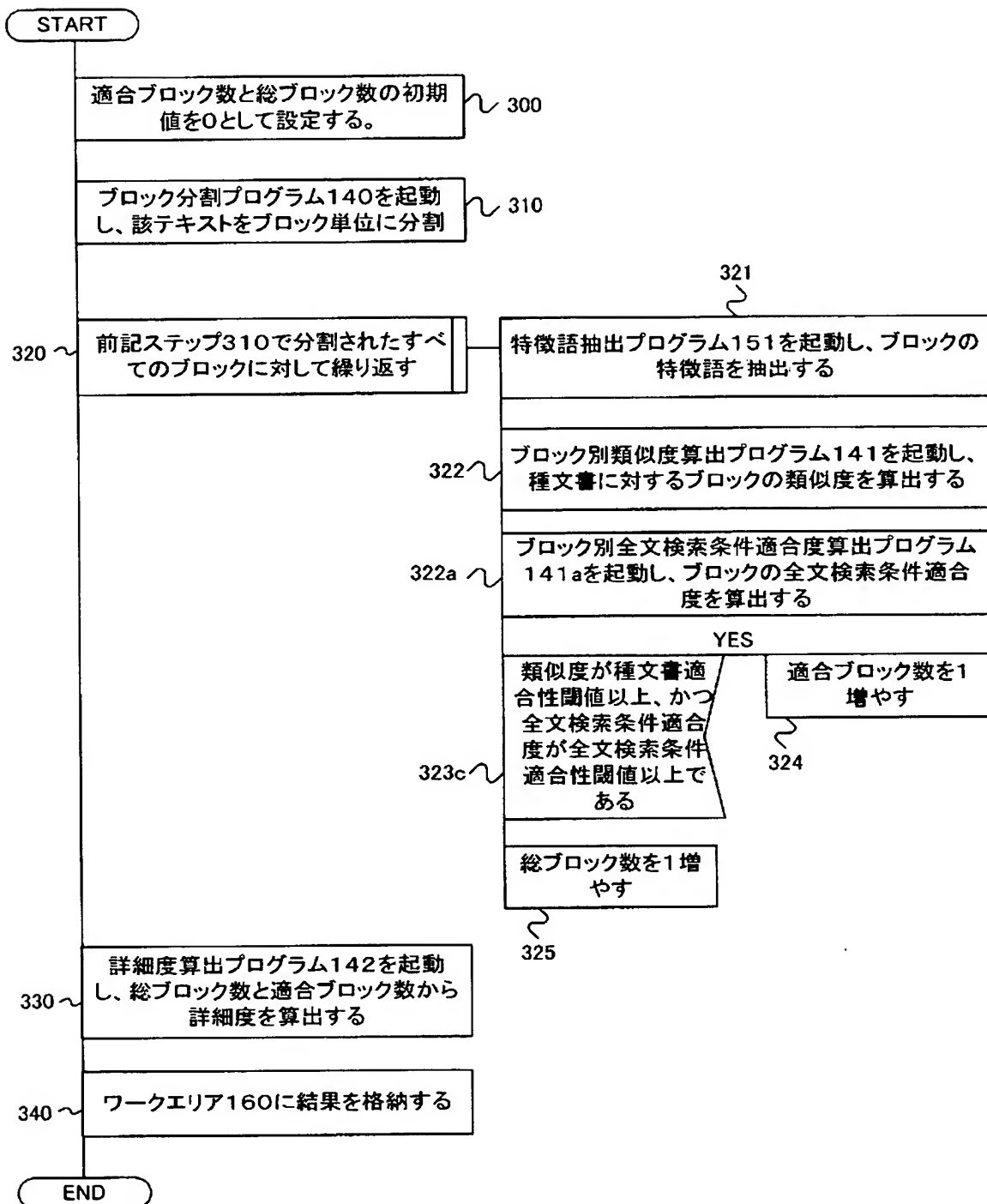
【図 10】

図10

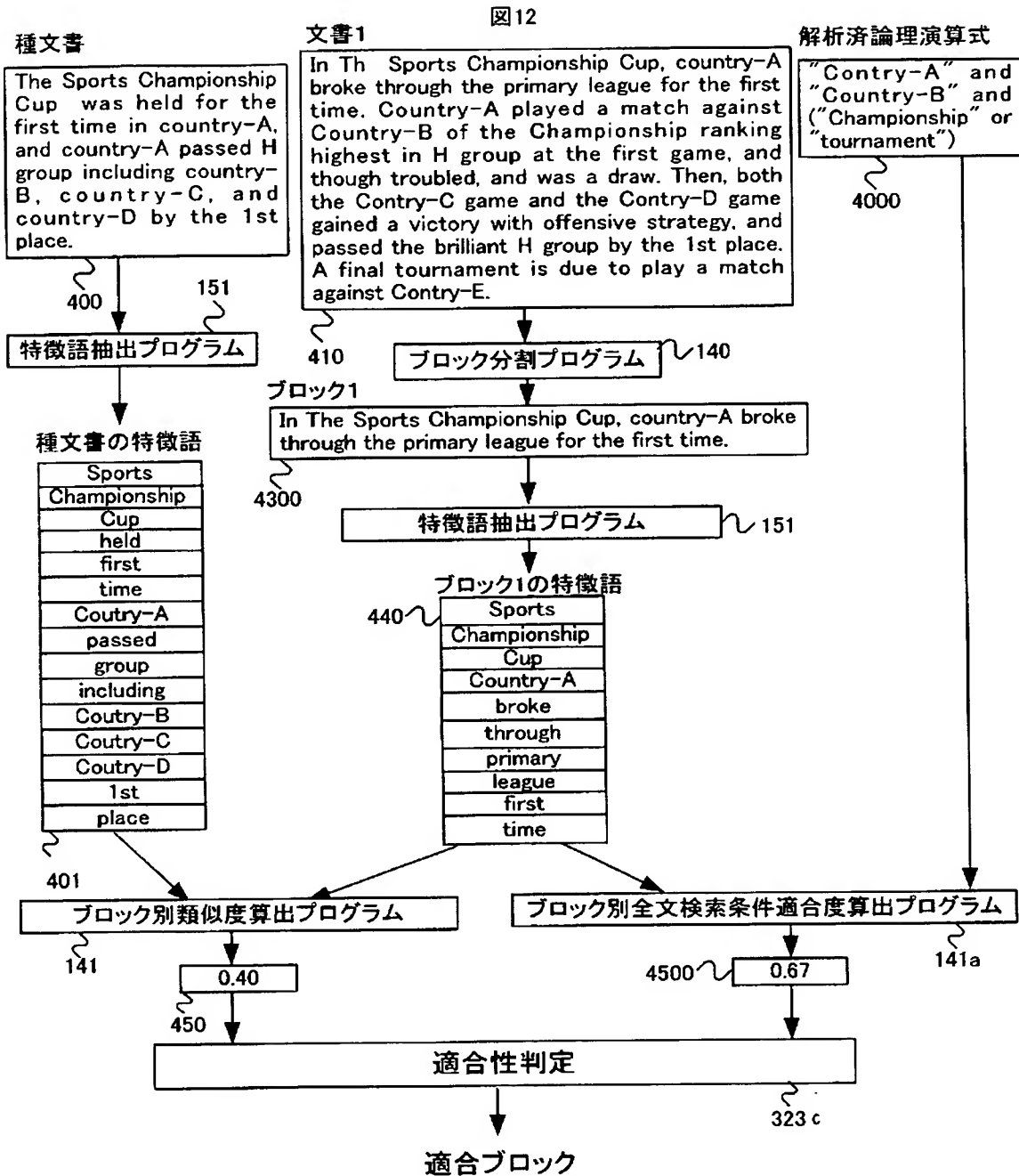


【図 11】

図 11

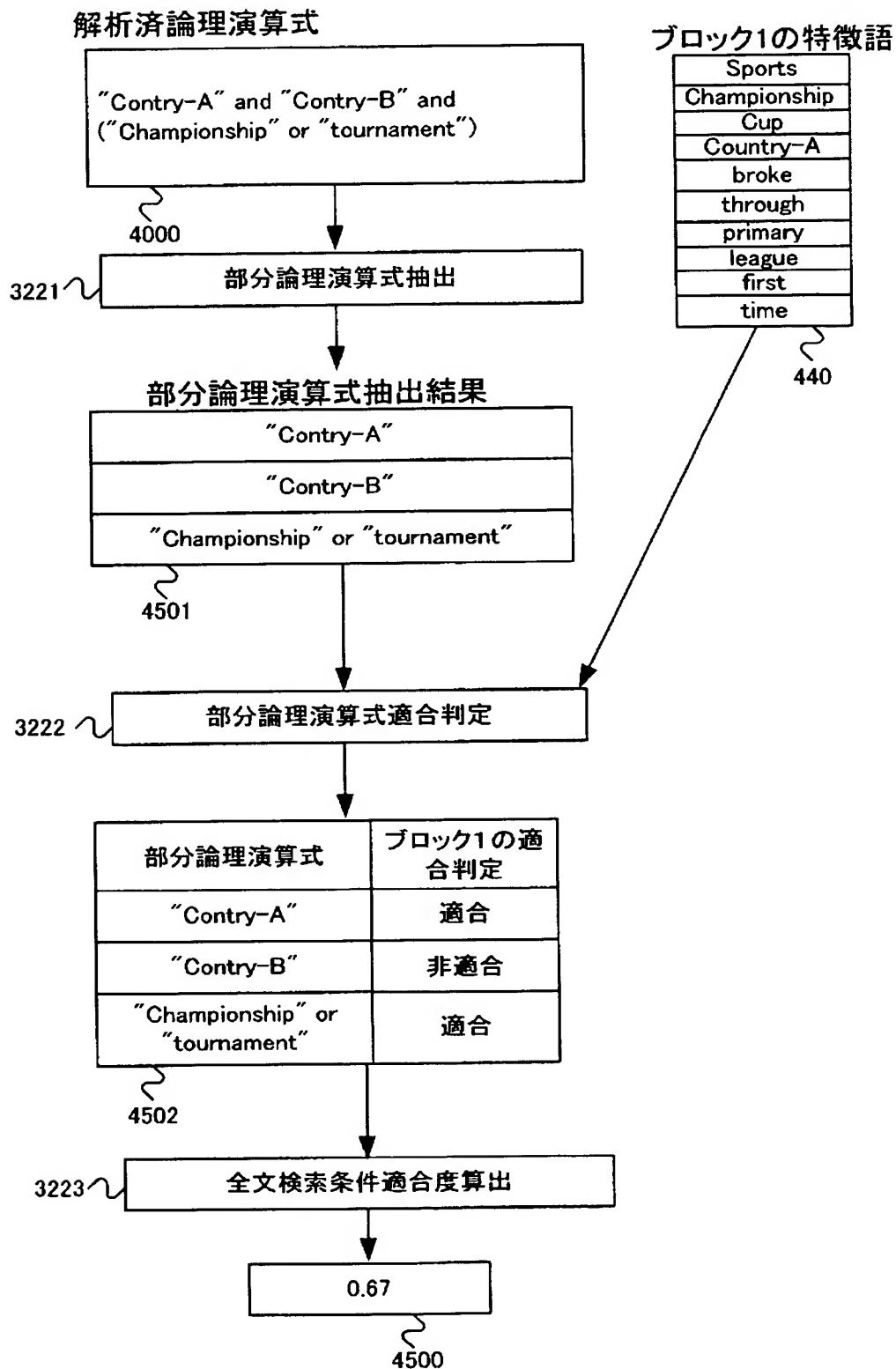


【図 12】

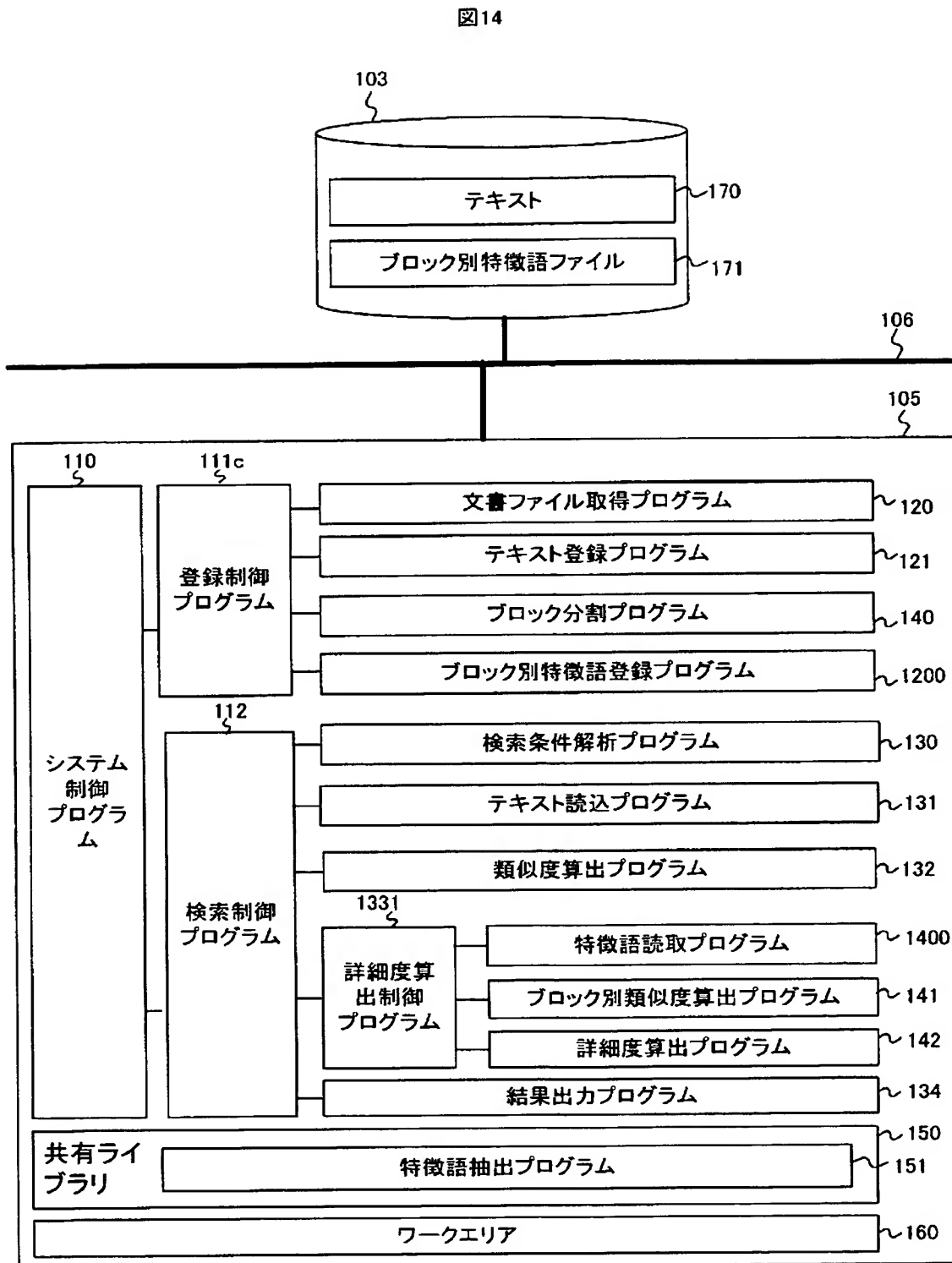


【図 13】

図13

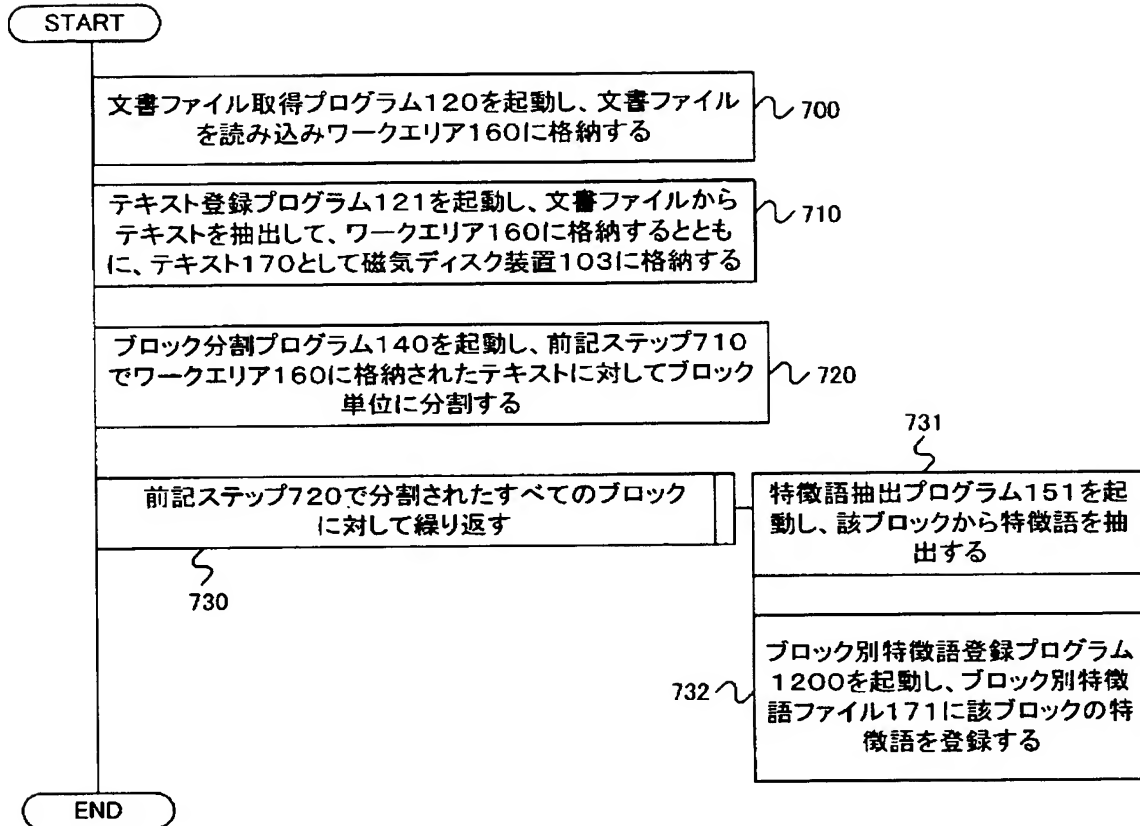


【図 14】



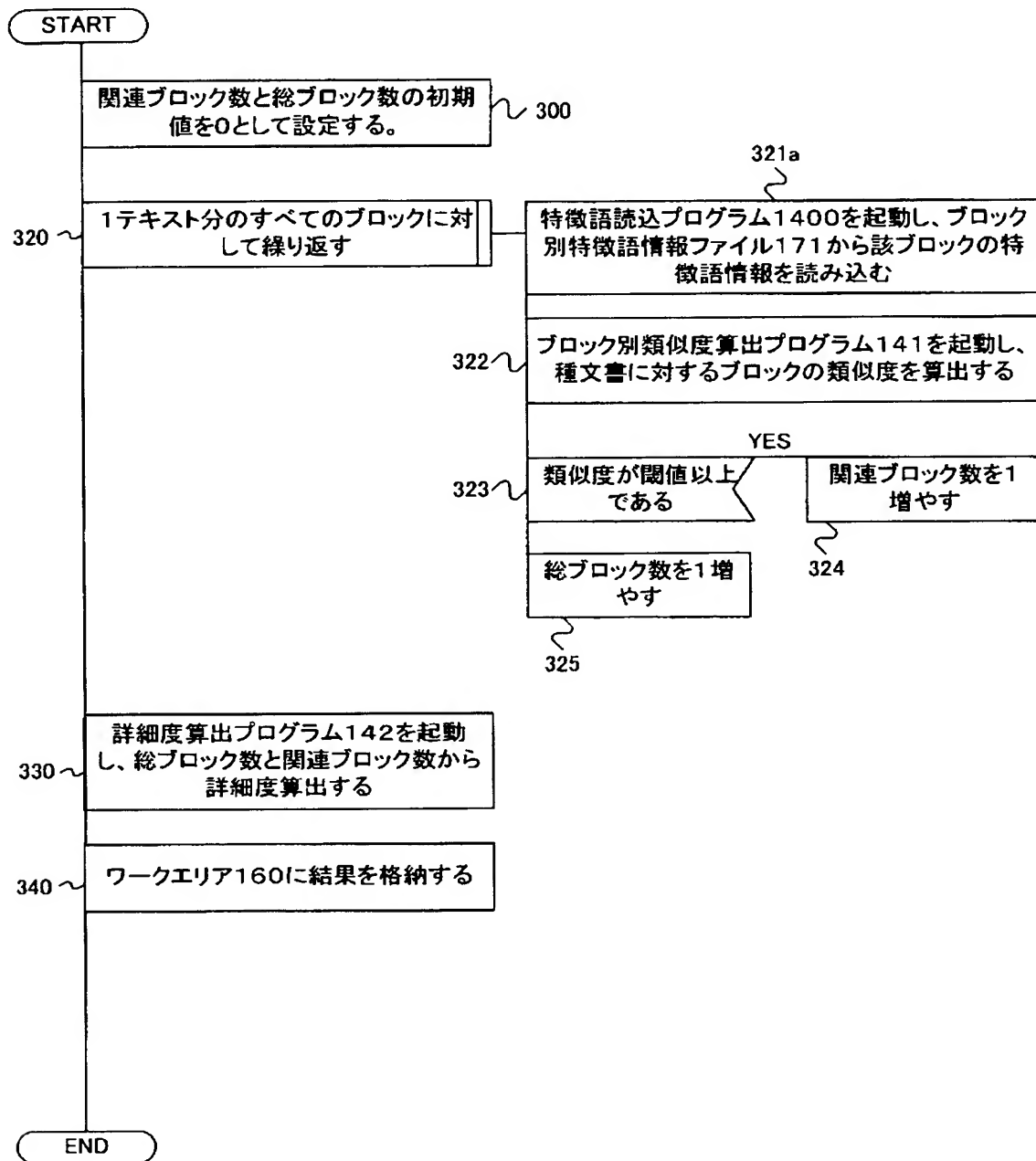
【図 15】

図15



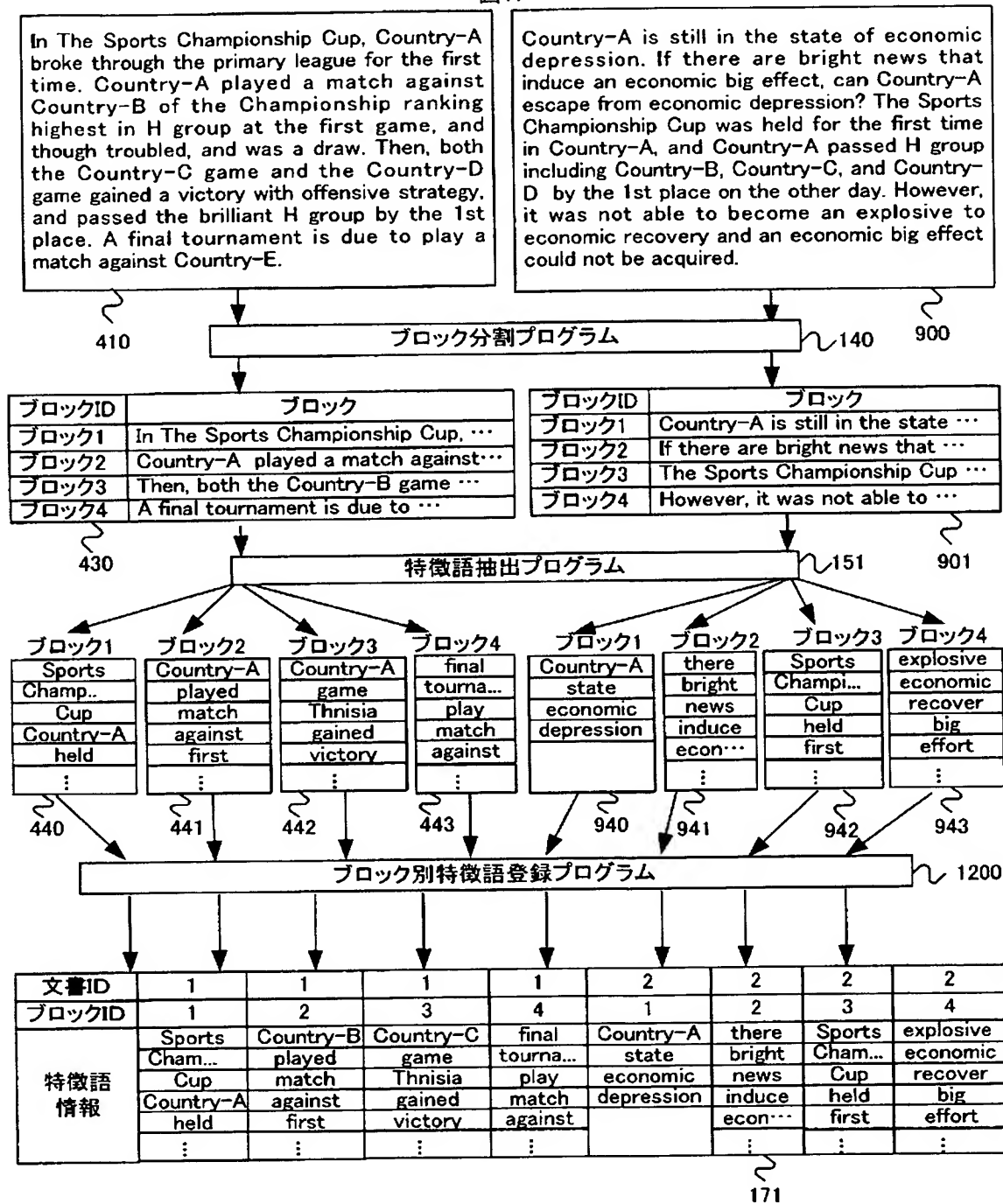
【図 16】

図16



【図 17】

図 17



【書類名】 要約書

【要約】

【課題】

文書間の類似性を判断するための指標を算出する類似文書検索方法を提供することにある。

【解決手段】

予め記憶された検索対象の対象文書の中から文書を検索する検索条件として入力された種文書に含まれる文字列を抽出し、対象文書を複数の部分に分割して、分割した対象文書の各部分に含まれる文字列を抽出し、これら文字列を比較して、前記分割された部分ごとに前記種文書に対する類似度を算出するとともに、その類似度と予め定められた閾値とを比較して、分割された各部分が種文書に適合している部分であるか否かの判定結果をもとに、対象文書の前記種文書に対する詳細度を算出する。

【選択図】 図 1

認定・付加情報

特許出願の番号	特願 2 0 0 3 - 0 8 9 6 3 3
受付番号	5 0 3 0 0 5 1 0 9 2 2
書類名	特許願
担当官	第七担当上席 0 0 9 6
作成日	平成 1 5 年 3 月 3 1 日

< 認定情報・付加情報 >

【提出日】 平成 15 年 3 月 28 日

次頁無

特願 2 0 0 3 - 0 8 9 6 3 3

出 願 人 履 歴 情 報

識別番号

[0 0 0 0 0 5 1 0 8]

1. 変更年月日

1 9 9 0 年 8 月 3 1 日

[変更理由]

新規登録

住 所

東京都千代田区神田駿河台 4 丁目 6 番地

氏 名

株式会社日立製作所